S177 Galaxies, stars and planets

The Open University

Galaxies, stars and planets

Science Short Module

The Open University

Galaxies, stars and planets

Simon Clark, Simon Green and Stephen Serjeant

This publication forms part of the Open University module S177 *Galaxies, stars and planets*. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 60 90; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858793; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

FSC
www.fsc.org
MIX
Paper from responsible sources
FSC® C013436

The Open University, Walton Hall, Milton Keynes MK7 6AA

First published 2012.

Edited and designed by The Open University.

Typeset by The Open University.

Printed in the United Kingdom by Latimer Trend and Company Ltd, Plymouth

ISBN 978 1 7800 7334 7

1.1

Chapter 1 The Universe today

1.1 Introduction

Astronomy is the study of all celestial bodies and the regions of space that separate them. It is a vast subject: quite literally as big as the Universe. It encompasses objects ranging in size from the incredibly small (the atoms from which planets and stars form) to the unbelievably vast (superclusters of thousands of galaxies, with each galaxy containing many billions of stars). The distances of these objects are also so enormous that the units of measurement, such as kilometres (km), that are familiar to many people on the Earth can't be used. The Universe is around 14 billion years old and contains materials with a range of properties that far exceeds anything that can be replicated in laboratories on the Earth. Temperatures range from close to absolute zero (−273.15 °C) in dense clouds of gas and dust from which stars form, up to many millions of degrees in the interiors of stars. Although interstellar space is not empty, its density is far lower than that found in the best vacuum chambers on the Earth. In some stellar remnants, material the size of a cube of sugar would have a mass of 100 million tonnes! How can anyone know so much about the Universe when no-one has ever explored or directly measured any of it except the tiny part very close to a rather ordinary star in one of many billions of galaxies, and for only a tiny fraction of the time the Universe has existed?

The 'scientific method' involves the observation or measurement of a phenomenon or property, followed by the formulation of a hypothesis (proposed idea) to explain it, which is testable by an experiment. The results of the experiment may confirm or support the hypothesis or may prove it to be incorrect, in both cases advancing our understanding and leading to further testable hypotheses. This objective approach allows different scientists to repeat experiments to confirm the results, and build on the knowledge and understanding obtained. It has underpinned the development of modern science. However, it cannot be applied to the study of most of the Universe because it's generally not possible to conduct experiments to test a hypothesis. It's only possible to observe the Universe and attempt to explain what can be seen. A hypothesis can only be tested by predicting what else might be seen if the hypothesis is valid, and then attempting to observe it. An observation which matches the prediction provides further evidence (but not necessarily proof) that the hypothesis is valid.

When astronomers make observations, this creates a snapshot of the Universe as it *appears* now. From this, astronomers must try to discover or deduce the properties of the different objects in the Universe, how they formed, and how they have evolved, over a 14 billion year time span! Many astronomical objects take millions or even billions of years to evolve. We cannot hope to record any process that takes more than a few thousand years to occur (the length of recorded history) or directly observe it for more than a few decades (the working lifetime of a scientist or the time since many modern astronomical methods were developed). Stellar evolutionary processes generally take far too long for anyone to see a noticeable change in their properties (although there are exceptions, as you will find out in Chapter 6). It's possible, however, to observe many different stars, all at different stages of their evolution, and try to build a picture of how they relate to each other in order to understand that evolution, in much the same way as attempting to work out the life cycle of a tree by walking through a forest.

The sheer scale of the Universe can help in this process. Objects can be seen from the light they emit. This light travels at a very high but finite speed of about three hundred thousand kilometres per second. The distances to stars are so vast that even at this speed it takes several years for light to reach us from the nearest stars. This is the origin of the use of 'light-years' as a measure of distance.

■ Why is a unit of time (years) used to measure distance?

□ Because a 'light-year' is the distance that light travels in one year.

Answering in-text questions

Throughout this book there are in-text questions (marked by a filled-in square), which are immediately followed by their answers (marked by a hollow square). To gain maximum benefit from these questions you should cover up the answer until you have thought of your own response. You will probably find it helpful to write down your answer, in note form at least, before reading the answer in the text.

The light from other galaxies can take many millions or even billions of years to reach us. This means that more distant galaxies can be seen as they were billions of years ago. More powerful telescopes, that allow us to see ever more distant objects, allow us to effectively observe further back in time and develop a better understanding of how the Universe and galaxies formed and evolved.

This first chapter summarises the scale of the Universe and the wide variety of objects it contains. Subsequent chapters will explain how the basic properties of the Universe are measured (Chapter 2), its history (Chapter 3), its limits (Chapter 4), its constituent stars (Chapter 5), how they evolve and how they change the composition of the Universe (Chapter 6) and the formation and properties of planetary systems (Chapter 7) as sites for the development of life (Chapter 8). A recurring theme will be how the composition of the Universe and the objects it contains has changed from its formation, through the lives of its constituent galaxies, stars and planets, through to the development of life (Figure 1.1).

1.2 Scale of the Universe

Our bodies, in common with all material on the Earth, the planets of our Solar System, the Sun, stars and galaxies are composed of what we call matter. Much of the matter in and among the stars is composed of atoms and

Figure 1.1 The image above illustrates some of the events that eventually led to life on Earth including the Big Bang, galaxies, the formation of stars and planets and the chemistry of life.

molecules that are familiar here on Earth. An atom is made up of particles called *electrons* which occupy a cloud-like structure around a central nucleus composed of *protons* and *neutrons*. The electrical attraction between negatively charged electrons and positively charged protons holds the atom together. A chemical *element* is defined by the number of protons in the nucleus. Hydrogen has one proton, helium has two and so on. A *molecule* is a group of two or more atoms (see Figure 1.2).

Many of the molecules in the human body are extremely complex combinations of atoms of different elements, but the simplest atom, hydrogen is the most common chemical element in the Universe, and helium is the second most common.

Atoms are among the smallest objects that are found in the Universe (individual hydrogen atoms are abundant in the interstellar medium – the space between the stars). Everyday units for length (such as metres) are just as

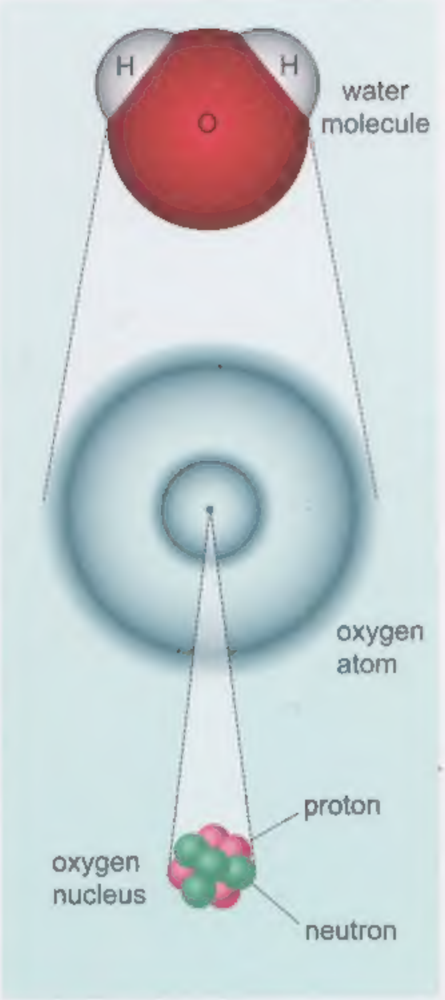


Figure 1.2 The structure of a water molecule which consists of two atoms of hydrogen and one of oxygen. The structure of the oxygen atom is also shown.

inappropriate for these tiny objects as they are for defining the vast distances between stars.

When measuring the scale of the Universe it's necessary to consider the smallest and largest objects and distances and how they're defined. The physical quantity of interest is length (it may be the diameter of an object or its distance from us).

If you would like some more practice with maths skills, see the *Maths Skills* ebook on the module website. Where you see the icon shown above, you can consult the relevant section of the *Maths Skills* ebook for further help.

Units, numbers and physical quantities

Much of astronomy concerns quantities such as distances, masses, temperatures, etc. In all of these cases, units of measurement are important. Physical quantities are generally the result of multiplying together a number and a unit of measurement. Thus a distance such as 5.2 metres is really the result of multiplying the number 5.2 by the unit of distance known as the metre. There are many units of measurement in common use, so, whenever you quote the value of a physical quantity, you should always take care to include the unit as well as the number multiplying that unit. It is no use being told that a distance is 5.2 if you don't know whether that means 5.2 centimetres or 5.2 metres. The unit is just as important as the number.

In scientific work there are several internationally agreed conventions for the definition of units and the way in which units should be used and represented when writing down the values of physical quantities. The most widely adopted system of units is known as SI, which stands for Système International. This is based on seven carefully defined units that include the metre (for length), the second (for time) and the kilogram (for mass). The other four base units relate to luminous intensity (i.e. brightness), quantity of matter, electric current and temperature.

The SI unit of temperature is called the kelvin (signified by the abbreviation K and not °K). A difference of 1 K is the same as a difference of 1 °C but the Kelvin scale starts from a different zero point. 0 °C corresponds to the temperature at which water freezes (at the Earth's atmospheric pressure). 0 K, which is equal to −273.15 °C, corresponds to *absolute zero* where, in the classical theory of matter, all motion of atoms ceases. Both scales are used in scientific literature. The Celsius scale, which is generally used in everyday life, is more commonly used for biological environments, whereas the Kelvin scale is used for the wider temperature ranges found in astronomy.

The recognised abbreviations for the metre, the second and the kilogram are m, s and kg, respectively. Abbreviated units should always be written in the singular form, i.e. 5.2 m, rather than 5.2 ms, since that might be misinterpreted as 5.2 × 1 m × 1 s or as 5.2 milliseconds, which is abbreviated to 5.2 ms. When writing units in full, for example as the result of a calculation, the singular should also be used (e.g. 5.2 metre rather than 5.2 metres). It is, however, acceptable to use the plural when expressing quantities in text to maintain correct grammar, as in the following paragraph.

In calculations, units should be treated in the same way as numbers. For example, speed is distance travelled divided by time taken, so the unit of speed is the unit of distance (metre) divided by the unit of time (second). For example, the result of dividing 6.0 m by 3.0 s is 2.0 m/s, which can be read as 2.0 metres per second.

Units that result from combining the *base units* are called *derived units*. The most common derived units are sometimes given their own names and symbols. Examples are joules (J; unit of energy) and watts (W; unit of power, which is energy per unit time or J/s).

The size of an atom is about one-ten-billionth of a metre, and the nucleus of that atom is around a hundred thousand times smaller still! Clearly some different units for distance are needed that are more appropriate for the very small and very large. One way to do this is to use units that are appropriate for a given scale. For example, the size of an atom is better described by a unit that is one-ten-billionth of a metre in size called an ångström (Å), whereas distances to stars can be measured in light-years (ly).

You have already been introduced to the light-year.

■ If light travels at 300 million metres in one second, how many metres are there in a light-year?

The number of seconds in a year is about (365.25 days) × (24 hours in a day) × (60 minutes in an hour) × (60 seconds in a minute), which equals over 30 million seconds (31 557 600 to be precise). So light will travel 300 000 000 × 31 557 600 = 9467 280 000 000 000 metres in a year.

You can see that 1 ly is easier to write than 9467 280 000 000 000 m! In fact professional astronomers use a different unit to measure the distance of stars, the parsec (pc), which is about 3.26 ly. You will learn more about the parsec in Chapter 2.

Ångströms, metres and parsecs cover an enormous range of sizes, and are ideal for measuring at atomic, human and stellar distance scales respectively. However, an even smaller unit than an ångström is needed to describe the sizes of atomic nuclei, and an even larger unit than the parsec for distances to other galaxies. Rather than invent new units every time we need to describe a certain scale, the SI unit of length, the metre, can be used. But how?

Very large and very small numbers

The SI unit of length is the metre. If you want to measure something that is typically 1000 times smaller or larger then you can use the familiar units the millimetre and kilometre respectively. So, using the same process, you could have units that are progressively 1000 times smaller or larger to reach any scale that's needed. The lists below show how:

Units larger than 1 metre:			
1 kilometre	1 km	is 1000 times larger than 1 m	
1 megametre	1 Mm	is 1000 times larger than 1 km	and 1000 000 × 1 metre
1 gigametre	1 Gm	is 1000 times larger than 1 Mm	and 1000 000 000 × 1 metre
1 terametre	1 Tm	is 1000 times larger than 1 Gm	and 1000 000 000 000 × 1 metre
1 petametre	1 Pm	is 1000 times larger than 1 Tm	and 1000 000 000 000 000 × 1 metre
1 exametre	1 Em	is 1000 times larger than 1 Pm	and 1000 000 000 000 000 000 × 1 metre
1 zetametre	1 Zm	is 1000 times larger than 1 Em	and 1000 000 000 000 000 000 000 × 1 metre
1 yotametre	1 Ym	is 1000 times larger than 1 Zm	and 1000 000 000 000 000 000 000 000 × 1 metre
Units smaller than 1 metre:			
1 millimetre	1 mm	is 1000 times smaller than 1 m	
1 micrometre	1 μm	is 1000 times smaller than 1 mm	and equals 1 m/1000 000
1 nanometre	1 nm	is 1000 times smaller than 1 μm	and equals 1 m/1000 000 000
1 picometre	1 pm	is 1000 times smaller than 1 nm	and equals 1 m/1000 000 000 000
1 femtometre	1 fm	is 1000 times smaller than 1 pm	and equals 1 m/1000 000 000 000 000
1 attometre	1 am	is 1000 times smaller than 1 fm	and equals 1 m/1000 000 000 000 000 000

So attometres would be useful for measuring atomic nuclei and petametres for measuring the distances to nearby stars. Figure 1.3 shows how these units relate to the scale of the Universe.

Most of these units will sound unfamiliar, not just because they refer to scales that we do not come across in our daily lives. Astronomy is an ancient science and certain non-SI units are commonly used instead, such as the ångström and parsec. Within our Solar System, distances are measured in units of the average distance of the Earth from the Sun, called the astronomical unit (AU). The rest of this chapter looks at the objects that can be observed in the Universe, starting with our own planetary system and the star that supports life on Earth, and then moving out to stars and galaxies.

1.3 Orbits and gravity

An understanding of orbital motion is fundamental to astronomy. It is crucial in the design of space missions and, as you will see in Section 7.6, it enables astronomers to deduce the existence of planets associated with other stars. Stars can orbit one another and they also move in orbit around the centre of a galaxy. This section introduces some key ideas about orbital motion.

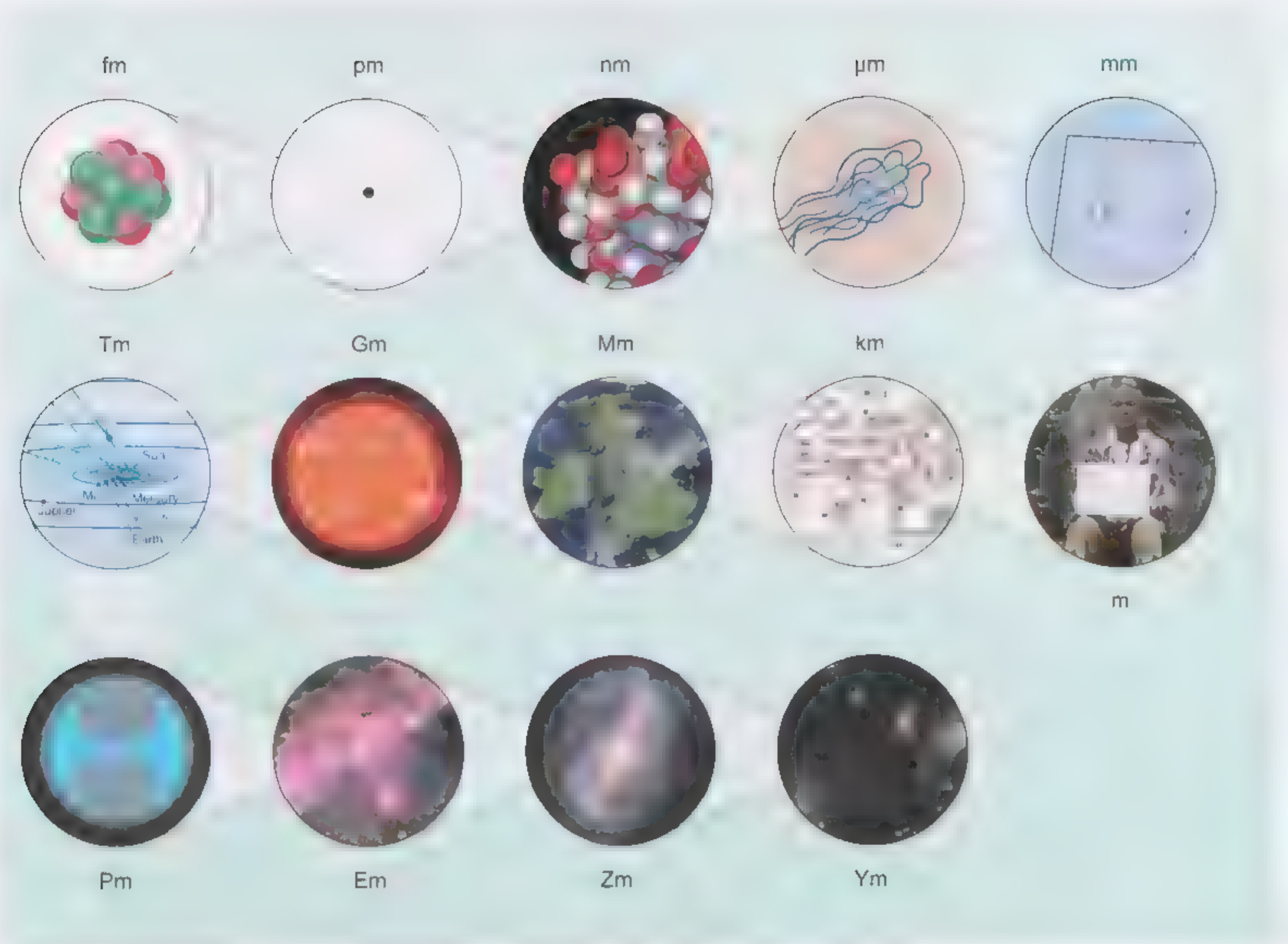


Figure 1.3 The scale of the Universe from atoms to galaxies. Each image is representative of the unit indicated. Each stage is 1000 times larger than the previous one. The smallest shown, femtometres (fm), is depicted by the nucleus of an oxygen atom. The next microscopic scale (pm) is still too small to show a whole atom. The next two scales (nm and μm) are represented by the diameter of deoxyribonucleic acid (DNA) and a bacterial cell respectively. You will be familiar with the scales millimetres (mm) to megametres (Mm). The Sun is at the scale of a gigametre (Gm). Moving outwards are the inner Solar System (Tm), Oort cloud (Pm), a nebula (Em), galaxy (Zm) and cluster of galaxies (Ym). You will meet these structures later in this book.

People sometimes wonder what keeps the Moon in orbit and stops it crashing to the Earth. This is a perfectly reasonable question to ask. If you lift an object above the Earth's surface and let it go, it falls to the ground, pulled down by the force of gravity. Why doesn't this happen to the Moon? Indeed, why don't the Earth and other planets fall into the Sun? To answer that question, and to see the role that gravity plays in the story, first requires an examination of circular motion.

If you set an object in motion it will move in a straight line, unless there is something pushing or pulling it into a curved path. To keep an object moving in a circular path, it needs constantly to be nudged sideways – there needs to be some force (that is, a push or a pull) directed towards the centre of the circle. The technical name for a force directed towards the centre of a circle is a **centripetal force** (centripetal means 'centre-seeking'). You can demonstrate this for yourself in the next activity.

Activity 1.1 Creating circular motion

The estimated time for this activity is 20 minutes.

For this activity you will need a table-tennis ball or large marble (or a similar smooth, smallish ball), a smooth table-top or floor, and about 1 metre of string (or wool) attached to a cork or a lump of plasticine (or other object of similar size and weight that can easily be fixed to your string). The second part of the activity (whirling the cork) needs to be done somewhere well away from people or objects that might be hit by a flying cork, ideally outdoors. Do not use a heavy object in this part of the activity.

First, roll the table-tennis ball (without spinning it) along the smooth surface. Note that it moves in a straight line.

Next, try to make it follow a curved path (again, without spinning it). You will find that, left to itself, it always follows a straight line. To get a curved path, you need to keep nudging it sideways, as shown in Figure 1.4. If you could exert a steady force rather than a series of taps, you could make the ball move in a smooth curve because you would be supplying the necessary centripetal force.

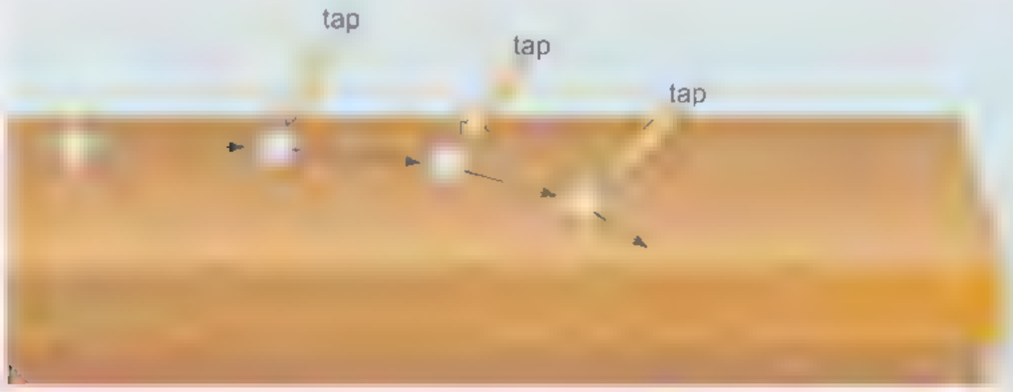


Figure 1.4 Making a table-tennis ball travel in a curve.

One way to supply a steady centripetal force is to pull on a piece of string attached to the moving object. Try whirling your cork or plasticine in a horizontal circle – you will feel that you need to keep pulling on the string as you do so.

Finally, let go of the string while whirling the cork and note the way it moves. You should be able to see that it continues to move in the direction it was heading at the time of release, as shown in Figure 1.5.

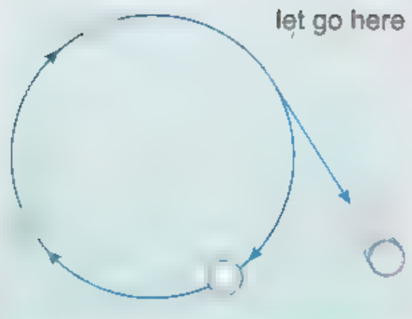


Figure 1.5 Letting go of a whirling object.

The need for a centripetal force applies to *all* cases of circular motion. In the example of the whirling cork, the force providing the inward pull is easy to see, but sometimes it is less so. For example, when a car is rounding a bend, the thrust of the engine and the grip of the tyres on the road combine to produce the necessary centripetal force.

What about the Moon? The Moon orbits the Earth in a (nearly) circular orbit.

- ■ What must be providing the centripetal force for the Moon's orbit?
- □ The force of gravity acting between the Earth and the Moon keeps the Moon moving in its nearly circular path.
- ■ What would happen to the Moon if gravity suddenly stopped acting?
- □ The Moon would drift off into space since there would be nothing to hold it in orbit around the Earth.

Gravity is familiar to us as the force that pulls objects towards the Earth, but our planet is not special in exerting this force. In fact, gravity acts between *all* objects. The strength of this attractive force increases in proportion to the ~~total mass~~ product of the masses of the two bodies and decreases in proportion to the square of the distance between the centres of the two bodies. Thus, the more massive the bodies and the closer they are, the stronger the force. Just as the Earth and the Moon are attracted towards each other by gravity, so too are all bodies. Even you and your cup of coffee are attracted to one another by gravity, but the force between small objects is so weak that it is normally unnoticed. We are usually only aware of gravity when at least one object is almost planet-sized.

Example: if you increase the distance by a factor of 3, the **force of gravity decreases by a** factor of three squared, written as 3^2 = 3 by 3 = 9. This will be described in more detail in Section 1.5.

If the Moon was simply suspended above the Earth and dropped, rather than moving in orbit, it would indeed move directly towards the Earth, pulled by gravity. In fact, it also has 'sideways' motion, and the overall effect is an orbit around the Earth. So, in wondering why the Moon does *not* fall towards the Earth, perhaps we should ask what gives it its sideways motion. Astronomers believe that the Moon formed from material ejected from the Earth in a giant impact. Some of this material would have ended up swirling around the Earth, where it gathered to form the Moon. The swirling motion is preserved in the form of the Moon's orbital motion.

So, in summary, it's possible to explain the Moon's orbital motion. It was acquired from the swirling material from which the Moon formed, and the Moon is kept in an almost circular orbit by the force of gravity acting between it and the Earth. Having explained the Moon's orbital motion, the same principles can be extended to other orbiting bodies.

1.4 Our neighbourhood

The **Solar System** consists of the Sun, eight major **planets**, some with one or more natural satellites and ring systems, and other minor bodies (**dwarf planets**, **asteroids** and **comets**).

Important terms are written in bold in this book.

Figure 1.6 shows the layout of the Solar System. All the planets orbit the Sun in the same **prograde** direction: anticlockwise when viewed from above the

North Pole. Their orbits lie roughly in the same plane and, except for Mercury, are almost circular. In Figure 1.6 the orbits are viewed from an oblique angle, which distorts their shapes. Orbits are discussed in more detail in Chapter 7.

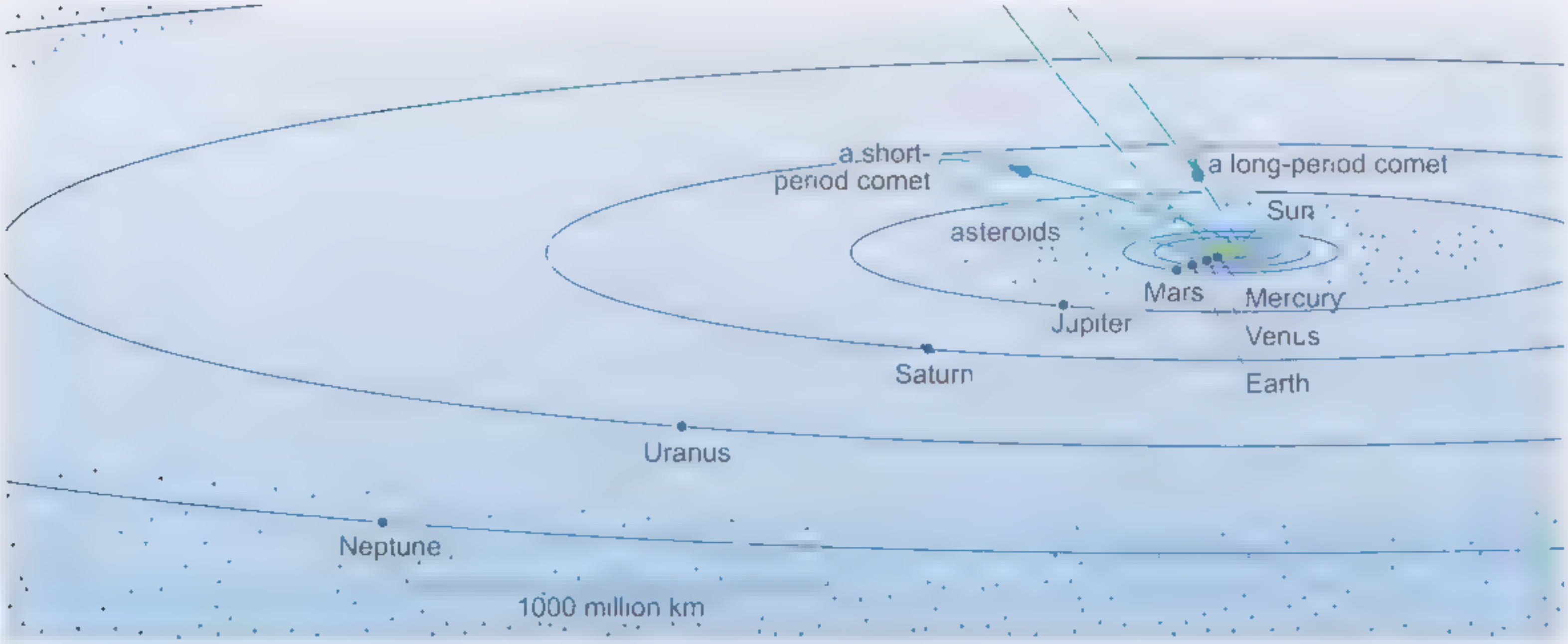


Figure 1.6 Schematic view of the Solar System showing the orbits of the eight major planets, looking obliquely southwards from outside the Solar System. Minor bodies are shown schematically. asteroids between Jupiter and Mars, trans-Neptunian objects in the outer Solar System, and the orbits of two typical comets. Note the scale bar showing a distance of 1000 million kilometres or 1 terametre.

Most of the planets spin on their axes with the same anticlockwise (prograde) sense of rotation. The exceptions are Venus, which spins very slowly backwards (**retrograde**), and Uranus, which is tipped on its side.

Figure 1.7 indicates the relative sizes of the major planets. Note the scale bar compared with that in Figure 1.6. The orbits of the planets cover distances of thousands of millions of kilometres, whereas Jupiter, the largest planet, is only 140 000 kilometres in diameter.

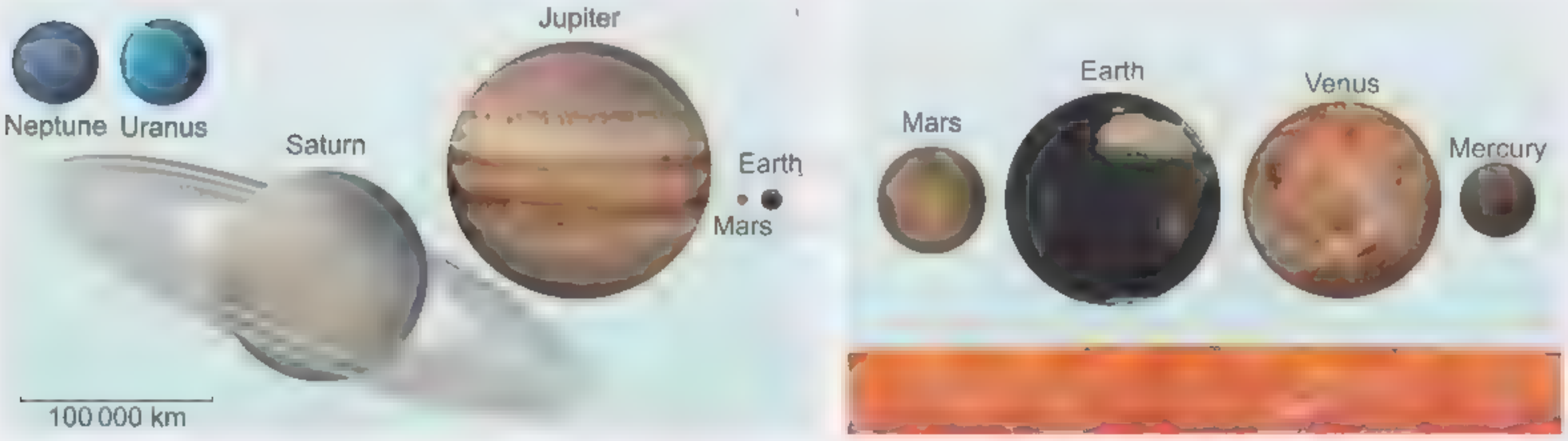


Figure 1.7 The Sun and the eight major planets showing their true relative sizes Note the change in scale between the right and left panels.

Table 1.1 lists the relative sizes of the planets on a scale of 1 cm to 5000 km. On this scale, a model Sun has a diameter of more than 2 m. You can get a feel for the relative sizes of the major bodies in the Solar System by representing each planet as a fruit. Note that an orange (Uranus) is about ten times the diameter of a redcurrant (Mercury), but the volume ratio is much larger – you could fit about 1000 redcurrants into the volume occupied by an orange.

Table 1.1 also shows the distances to the planets on the same scale. If you made a scale model of the Solar System you would not want to arrange the model planets at these relative distances! This illustrates the vast distances between the planets compared with their sizes.

Table 1.1 The sizes and distances of the planets on a scale of 1 cm to 5000 km.

Planet	Approx. diameter/ km	Approx. model diameter/ cm	Representative fruit	Approx. distance from Sun/million km	Approx. model distance/m
Mercury	5 000	1.0	redcurrant	58	116
Venus	12 000	2.4	cherry tomato	108	216
Earth	13 000	2.6	cherry tomato	150	300
Mars	7 000	1.4	blueberry	228	456
Jupiter	140 000	28	water melon	778	1600
Saturn	120 000	24	pumpkin	1430	2900
Uranus	51 000	10	orange	2870	5700
Neptune	49 000	9.8	orange	4500	9000

■ The data in Table 1.1 are for a model on a scale of 1 cm to 5000 km. How much bigger is the real Solar System than the model?

One metre is 100 centimetres, and 1 kilometre is 1000 metres, so there are one hundred thousand centimetres in a kilometre (i.e. 1 km = 100 000 cm). In 5000 kilometres there are five hundred thousand thousand centimetres – in other words five hundred million centimetres – so the actual Solar System is five hundred million (500 000 000) times bigger than the model.

The planets form two groups: the four closest to the Sun (the **terrestrial planets**) are similar in size to the Earth and have rocky surfaces, whereas the outer four planets (**gas giants**) are much larger with deep dense atmospheres. The reasons for these differences will become apparent in Chapter 7.

The orbits of all planetary satellites lie close to the plane of their planet's equator and most travel in the same prograde direction as their planet's spin The largest are comparable in size with the planet Mercury, whereas the smallest are little more than giant boulders. The largest of the minor bodies (**asteroids**, **comets** and **trans-Neptunian objects**) are more than 1000 km in diameter and are large enough to have their shapes (roughly spherical) determined by their own gravity – they are called *dwarf planets* and include the former planet Pluto as well as the largest asteroid Ceres.

For astronomers, the Sun is fascinating because it is our nearest **star**. By studying the Sun, they can gain an insight into the workings of the other millions of stars that are visible in the night sky. Learning that the Sun is a star can be a little surprising. After all, the Sun is a brightly glowing, yellow object – so bright that it is dangerous to look at it directly, and so hot that its radiation can be felt warming the whole Earth. Stars, on the other hand, are mere pinpoints of light that are visible only against the darkness of the night sky and with no discernible heating effect on Earth. How can they possibly be the same sort of object? The key to the answer lies in their *distances*.

In astronomical terms, the Sun is relatively close, being only about 150 million kilometres (1 astronomical unit) from Earth. As you have seen, the stars that are visible at night are so much further away that they appear as just faint points of light. Imagine looking at a glowing light bulb first from very close up and then from a much greater distance. Close up, you would see the shape of the bulb but, from far away, it would be just a point of light.

Although it is a very ordinary star, the Sun dominates the Solar System. With a diameter of 1.4 million km it is about ten times larger than Jupiter and more than a hundred times larger than the Earth. Its mass is over three hundred thousand times that of the Earth. The combined mass of the planets is less than 0.2% of the mass of the Sun. For this reason the Sun dominates the Solar System in several ways. The Sun's gravitational force controls the motion of bodies within the Solar System (see Chapter 7). It also distinguishes the Sun (as a star) from planets. The temperatures and pressures near the centre of this massive body are sufficiently high to sustain the nuclear reactions that power the Sun and result in its prodigious output of energy in the form of electromagnetic radiation (you will learn more about the radiation from the Sun in Chapter 2 and nuclear reactions in stars in Chapter 5). The planets, with their much smaller masses, cannot support these reactions. They are generally observed as a result of reflected or absorbed and re-emitted sunlight.

1.5 The Sun and stars

Safety warning

Never look directly at the Sun, either with the unaided eye or through spectacles, binoculars or a telescope. You risk permanently damaging your eyes if you do so.

The part of the Sun that you normally see is called the **photosphere** (meaning 'sphere of light'); this is best thought of as the 'surface' of the Sun, although it is very different from the surface of a planet such as Earth. The photosphere is not solid. Rather, it is a thin layer of hot gaseous material, about 500 kilometres deep, with an average temperature of about 5500 °C or 5800 K.

Figure 1.8a shows a visible light image of the Sun. **Sunspots** (Figure 1.8b) appear as small dark patches on the photosphere, which typically survive for a week or so, and sometimes for many weeks. The longer-lasting sunspots can be photographed repeatedly as they cross the face of the Sun so they can even be used to investigate the rate at which the Sun rotates. The number of sunspots changes with time, gradually increasing to a maximum every eleven years then decreasing to a minimum when no sunspots may be visible for some time — a cycle intimately linked to changes in the Sun's magnetic field. A close-up view of the visible surface of the Sun (Figure 1.8b) also reveals a seething pattern of **granules** seen all across the photosphere. Individual granules, resulting from upwelling hot gas due to convection (an important process in stellar energy transport described in Chapter 5) come and go in a few minutes, often to be replaced by other granules.

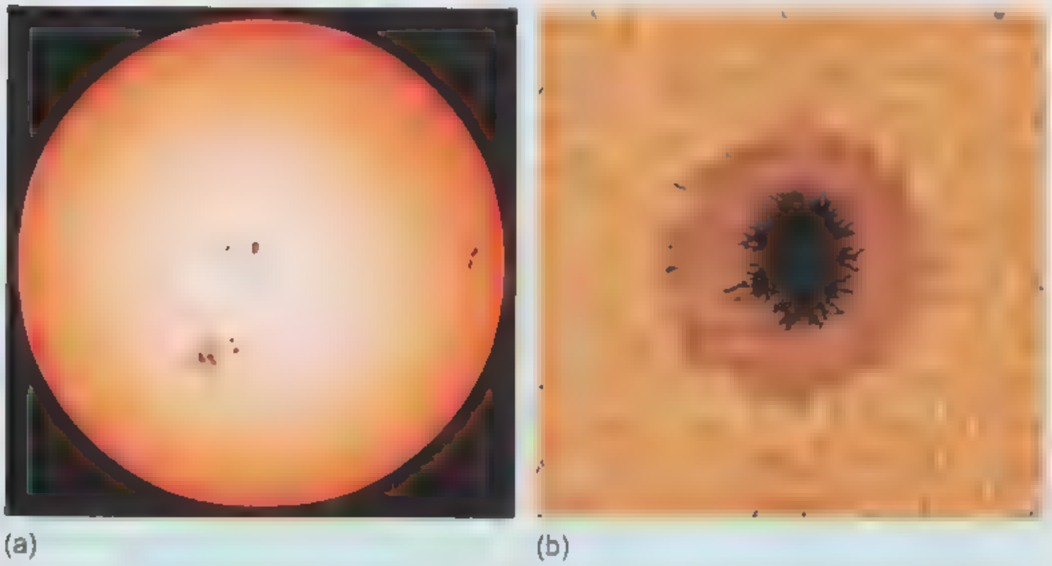


Figure 1.8 (a) A visible light image of the Sun. (b) A sunspot and granules on the Sun's surface.

Detailed studies of the body of the Sun usually require special equipment. However, the natural phenomenon known as a **total eclipse of the Sun** provides an opportunity to gain further insight into the nature of the Sun (see Figure 1.9). A total eclipse happens when the Moon passes in front of the Sun and blocks out the bright light from the photosphere.

When the Moon just eclipses the bright photosphere, it is often possible to see part of a narrow, pink-coloured ring that encircles the Sun. This is the **chromosphere** (meaning 'sphere of colour'), the lower or 'inner' part of the Sun's atmosphere. It is actually another layer of gaseous material, a few thousand kilometres thick, which sits on top of the photosphere. The lower parts of the chromosphere are cooler than the photosphere, while the higher parts are considerably hotter, but the chromospheric material is so thin that it emits relatively little light, and is therefore unseen under normal conditions.

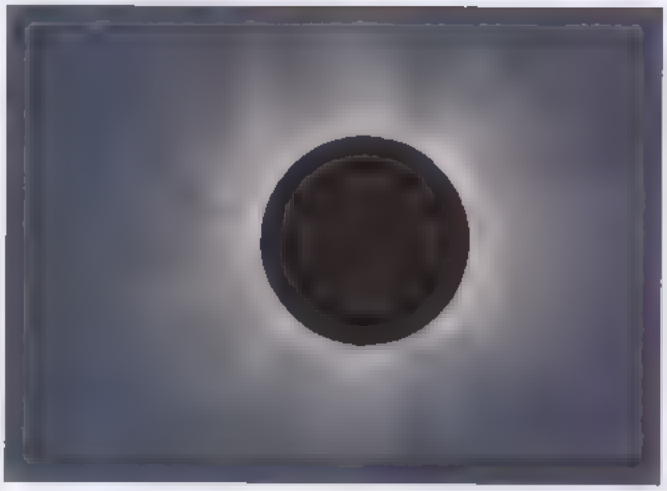


Figure 1.9 A total eclipse of the Sun, revealing the inner and outer parts of the Sun's atmosphere, the chromosphere and the corona.

As a total solar eclipse proceeds, a third part of the Sun is seen the **corona** (meaning 'crown'). This is the extremely tenuous (i.e. thin) upper atmosphere of the Sun that extends to several times the Sun's photospheric radius. The corona sometimes looks like streamers or plumes, but its shape changes from eclipse to eclipse, although it will not usually show any changes during the few minutes of totality that characterise a typical total eclipse. The corona is very hot (temperatures of several million kelvin are not unusual) but it is so thin that its pearly white light is very faint compared with the light from the photosphere.

■ The corona may be faint, but it does glow. Why are we not normally aware of the Sun's corona?

The bright light from the Sun's photosphere is scattered by the Earth's atmosphere. This makes the sky blue and generally rather bright. As a result, the much fainter light from the corona can't be seen (rather as the light from a dim torch is unnoticeable on a bright sunny day).

Sometimes in eclipses observers also see **prominences** great spurts of hot material at the edge of the Sun, extending outwards from the solar surface for many thousands of kilometres. Prominences and the changing shape of the corona indicate that the Sun is an active body, not just a quietly glowing source of light. When the Sun is observed with instruments that can detect electromagnetic radiation other than visible light, it is possible to see the full extent of the activity of the Sun. You will learn more about this in Chapter 2.

Prominences, sunspots and other features of the Sun seen at different wavelengths are indicative of **active regions**, generally caused by the Sun's magnetic field, which influences the flow of hot gaseous material on the Sun. Sudden changes to the magnetic field in the corona are thought to be responsible for **flares**, one of the most energetic of all solar phenomena, which emit bursts of radiation of all wavelengths, from radio waves to gamma-rays as well as energetic particles (such as fast-moving protons and electrons).

The Sun is a typical star and only appears much brighter than other stars because they are so much further away. Astronomers can use this to deduce the actual distances to stars. One important observation makes this easier: namely, *stars that are the same size and colour give out the same amount of light*. So, if astronomers observe two stars of exactly the same colour, they can start by *assuming* they are the same size and therefore they must be giving out the same amount of light. If one looks fainter, it must be further away. By measuring the amount of light entering a telescope from each star, astronomers can work out just how much further away one star is than the other. Figure 1.10 shows the principle. Stars A and B give out the same amount of light, but B is at twice the distance of A, so its light is more spread out by the time it reaches an observer on Earth. Four (2 × 2) times as much light from A enters the telescope (or eye), so A appears four times brighter.

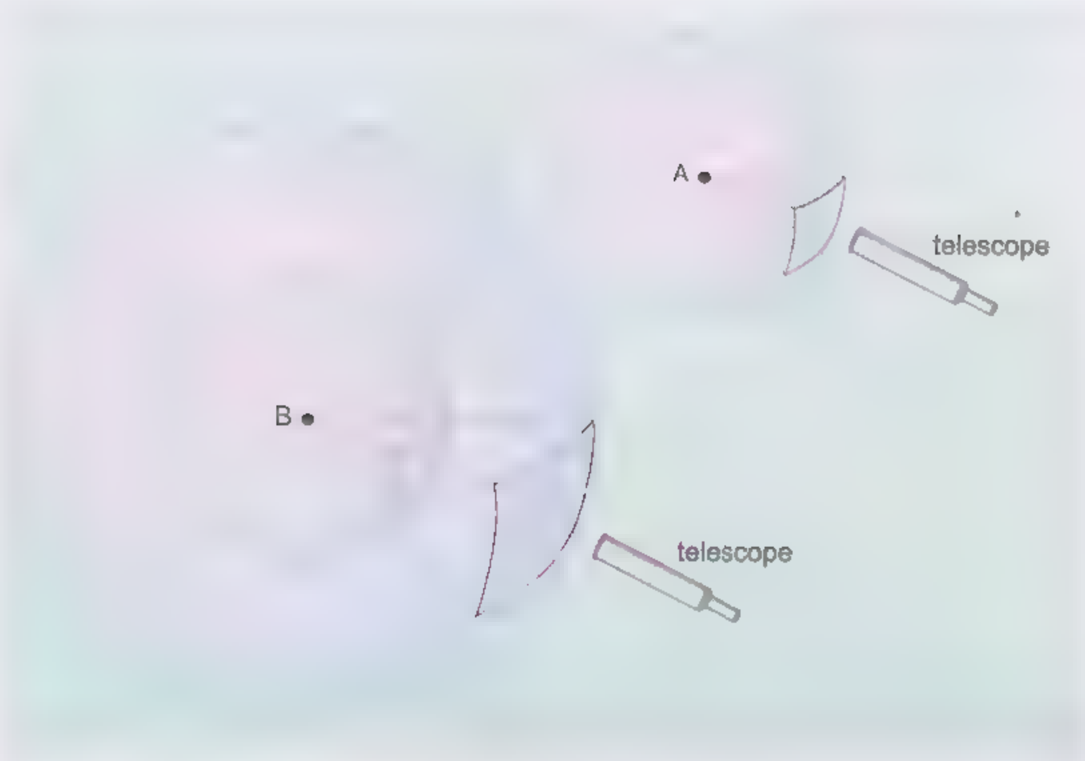


Figure 1.10 Light from a more distant star B is more spread out, so the star appears fainter than an identical star A nearby.

Powers

When a number is multiplied by itself, the result is called the *square* of that number. For example, multiplying three by three gives nine. Nine is said to be 'the square of three' or 'three squared'. This is expressed mathematically as 3^2, which means 3 × 3. The *cube* of a number is that number multiplied by itself three times, so three cubed = $3 \times 3 \times 3 = 3^3$. Note that 3^3 can also be described as 'three to the power of three'.

■ What is the general rule for describing how the apparent brightness of a star diminishes with distance?

□ The general rule is that the apparent brightness diminishes with the *square* of the distance.

So if the distance is multiplied by 2, the apparent brightness is reduced by $2^2 = 2 \times 2 = 4$. If the distance is multiplied by 5, the apparent brightness is reduced by $5^2 = 5 \times 5 = 25$, i.e. such a star has 1/25 of the apparent brightness of a similar star lying at 1/5 of its distance.

Squares can also be worked out backwards so that as nine is three squared, three is the *square root* of nine, written $3 = \sqrt{9} = 9^{1/2}$, and as 27 is three cubed, three is the *cube root* of 27, thus $3 = \sqrt[3]{27} = 27^{1/3}$.

The apparent brightness of a star (or any other luminous object) is therefore said to obey an **inverse square law**.

- If star B is 10 times the distance of identical star A, how much brighter would A appear?

  Star A would appear 10 × 10 = 100 times brighter than star B.

In practice, it is not quite so easy to measure distance, because some stars are the same colour but different sizes and so give out different amounts of light – but the general principle of 'faint means far' underlies many of the techniques for measuring distances. The determination of distances will be discussed further in Section 2.5. Figure 1.11 shows that stars have different colours These colours are related to the temperatures of the stars. The Sun is yellowish, with a photospheric temperature of about 5800 K. Bluish white stars are hotter than the Sun and orange–red stars are cooler. Stellar temperatures range from less than 2000 K to over 40 000 K In Section 2.3 you will learn more about how temperatures of stars are derived.

Figure 1.11 A region of sky visible from the Southern Hemisphere. The brightest star is Alpha Centauri, a faint companion of which is the closest star to our Sun. To the right is the famous constellation of the Southern Cross (Crux).

Stars come in a range of sizes, masses and luminosities. **White dwarfs** are only around the size of the Earth whereas some **red giants** are so large that if placed at the location of the Sun, they would engulf the Earth! The masses of stars, however, cover a much smaller range. The least massive are around ten percent of the mass of the Sun and the most massive around a hundred solar masses. The mass of the Sun, 1 solar mass, denoted $1\,M_{\odot}$, provides another commonly used unit in astronomy. The subscript symbol ⊙ represents the Sun, so the radius of the Sun is $1\,R_{\odot}$ and the luminosity (the total power output) of the Sun is $1\,L_{\odot}$. The luminosities of stars range from less than a thousandth of the solar luminosity to greater than 1 million $L_{\odot}$. Figure 1.12 shows examples of different types of stars. A wide variety of combinations of properties are found, such as small, cool, faint red dwarfs and large, hot, highly luminous supergiants. However, some types are more common than others and not all possible combinations of these properties are found among the stars that have been observed. You will find out in Chapters 5 and 6 why this is the case and what can be learnt about the evolution of stars from these properties.

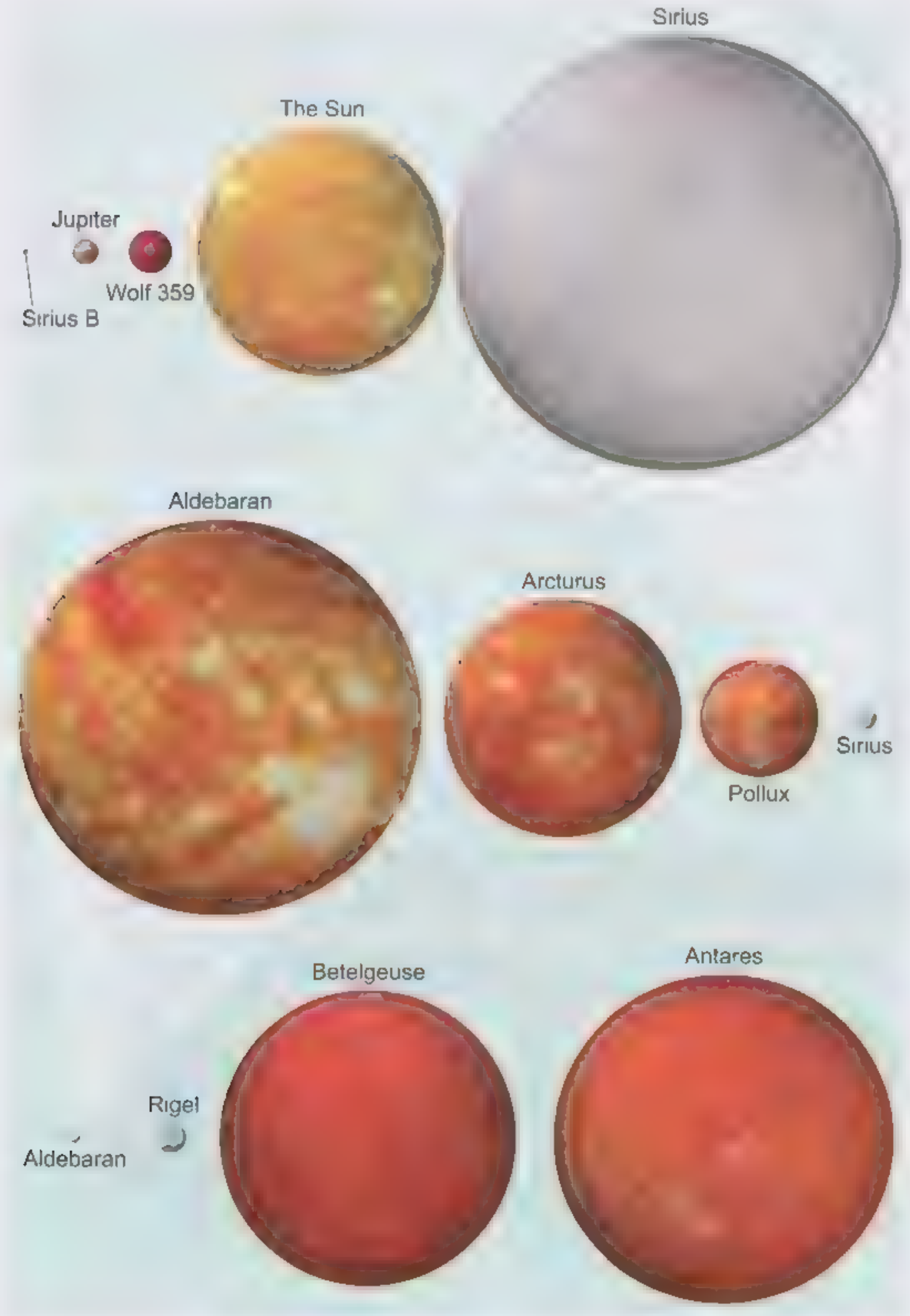


Figure 1.12 Sizes of stars compared. The colours are indicative of temperature; the bluish stars are hotter than the Sun whereas the orange red stars are cooler. Jupiter, shown in the first panel, has too small a mass to support the nuclear reactions that power the stars.

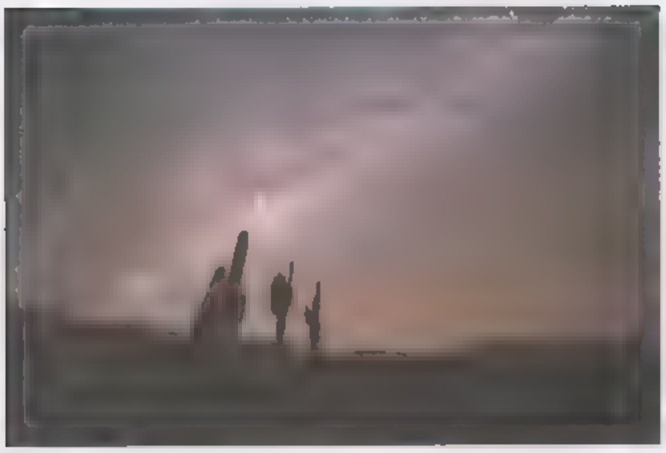


Figure 1.13 The night sky as seen from the Atacama Desert in Chile, showing the Milky Way in the direction of the centre of the Galaxy.

1.6 Galaxies

The Sun is one of about a hundred billion (10^{11}) stars in our galaxy. It is difficult to determine the structure of the Galaxy as we are located inside it. As well as stars, it contains vast clouds of gas and dust, which can obscure our view in certain directions. However, if you observe the night sky from a dark site on a clear moonless night, you will see the Milky Way, a band of light circling the sky, that comes from many faint stars that cannot be individually distinguished (Figure 1.13). It reveals the most obvious characteristic of our galaxy, that it has a flattened shape. Careful analysis of the distances and motions of stars in space (see Chapter 2) are required to reveal the true nature of our galaxy, called the Milky Way galaxy

Our galaxy is sometimes referred to as 'the Galaxy' (with a capital 'G') to distinguish it

The human eye is extremely sensitive, but even with the aid of a telescope it is not ideal as an astronomical detector, because it does not record images Until the use of photography in the late nineteenth century, astronomers recorded their observations with drawings made at the telescope Despite the fact that photographic plates were much less sensitive than the human eye, they had one additional critical advantage they could accumulate the light from a faint object for as long as a telescope could track it (the human eye retains an image for less than a tenth of a second). Photographs, and more recently electronic imaging detectors, reveal a huge variety of galaxies (Figure 1.14). They revealed that what appeared as faint smudges of light to the eye were in fact vast systems of stars like our own galaxy. You will learn about the different types of galaxy, how their appearance is related to their structure and how they have evolved throughout the history of the Universe in Chapter 3.

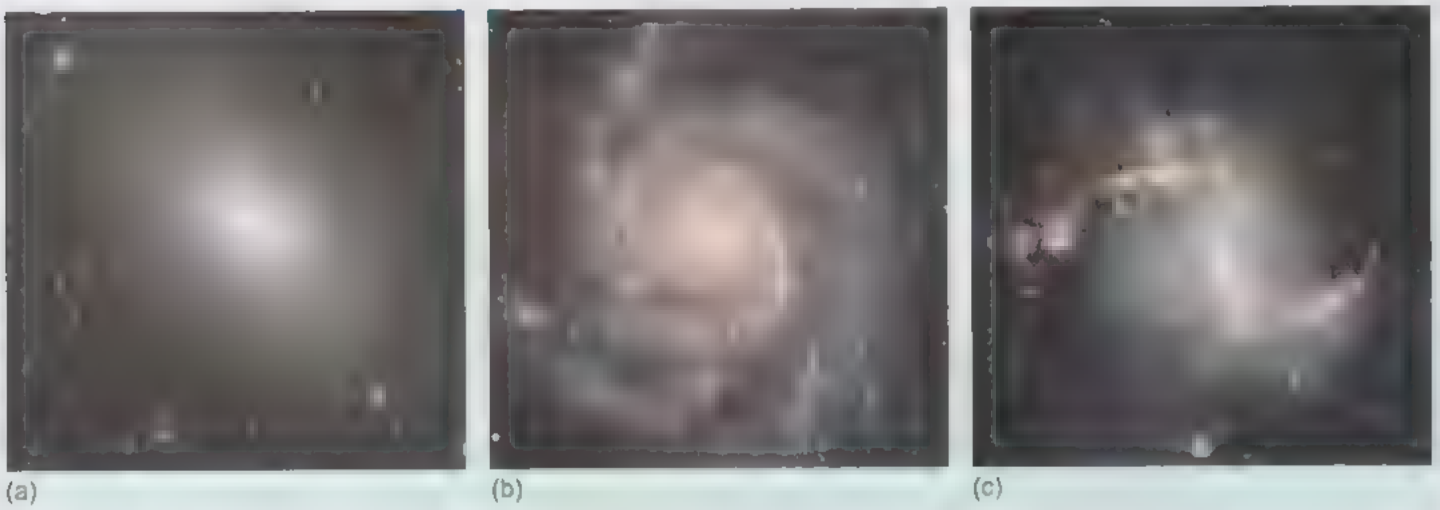


Figure 1.14 Examples of different types of galaxy. (a) An elliptical galaxy. (b) A spiral galaxy (c) An irregular galaxy.

Galaxies are not distributed uniformly in space. Our own galaxy is a member of a small **group** of about 40 galaxies. Larger **galaxy clusters** may have more than a thousand members (see Figure 1.15) and these clusters themselves appear to be arranged into even larger structures.

Our understanding of the Universe is, not surprisingly, derived largely from the light emitted by stars and galaxies. However, our understanding of the properties and evolution of these stars and galaxies comes from applying scientific principles and mathematical models. As new observational techniques developed and this understanding grew it became more apparent that the objects that can be seen represent only a fraction of the matter in the Universe. The majority of matter, called **dark matter,** is not visible but is required to understand the properties of the Universe. Some dark matter may simply be in the form of dead stars, but most appears not to be made up of the familiar elements but of some so far unknown constituents. This material is called **non-baryonic dark matter**. The evidence for dark matter, its types and possible nature will be discussed in Chapters 3 and 4. Figure 1.16 indicates what is currently thought to be the material composition of the Universe.

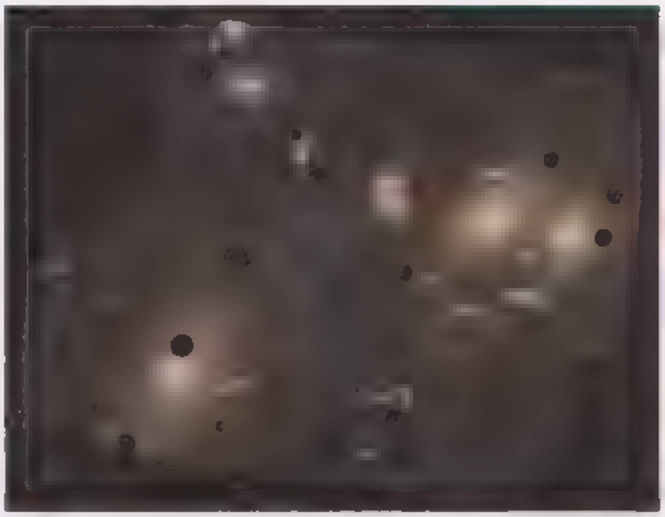


Figure 1.15 The Virgo Cluster of galaxies. The dark spots indicate where bright foreground stars were removed from the image. Messier 87, which is visible through a small telescope, is the largest galaxy in the picture (lower left).

1.7 'We are stardust'

How did the atoms in our bodies, the Earth, the Sun and other astronomical objects originate? This question is intimately tied with one of the most fundamental questions in science 'How was the Universe formed?'. As you will discover in Chapter 4, only the simplest atoms were present in the early stages of the Universe. Even now, the composition of normal matter in the Universe is dominated by hydrogen and helium. All the other heavier atoms, rather confusingly referred to as 'metals' by astronomers, amount to around 2% of normal matter. The heavier elements have been produced in nuclear reactions within stars; you will learn more about these reactions in Chapter 6.

If the nuclear reactions in stars occur close to the centre where the temperatures and pressures are highest, how can these heavy elements escape to make their way into the gas and dust clouds in the Galaxy and ultimately into planets and us? The answer lies in the details of the evolution of stars. The structure of stars changes as they age and some end their lives in catastrophic explosions that distribute much of their mass into the surrounding space and increase the fraction of heavier elements that can be incorporated into new stars.

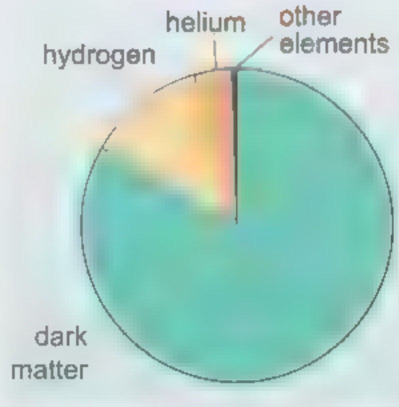


Figure 1.16 The matter in the Universe Chemical elements account for less than about one-sixth of all matter. The majority is believed to consist of non-luminous 'dark matter', the nature of which is still uncertain.

- About 1% of the Sun's photosphere is composed of heavy elements. If the Sun formed from clouds of gas and dust in the Galaxy why does it contain only half the amount of heavy elements now present in the Galaxy?

  The Sun formed when the Galaxy was considerably younger (around 4.6 billion years ago). As time passes more and more stars complete their life cycles and so the proportion of heavy elements increases The Sun therefore formed when there were fewer heavy elements in the Galaxy.

The evolution of stars and the ways in which they enrich the interstellar medium are described in Chapters 5 and 6.

Table 1.2 lists the fraction by mass of the most common elements in the photosphere of the Sun compared with those in the Earth and a human body. As you would expect, the elements that are most common in the Sun are those that are most commonly produced in the nuclear reactions in stars. However, if the Earth formed at around the same time as the Sun why does it have such a different composition? The most abundant atoms in the Universe are only minor constituents of the Earth. The answer lies in the way in which these elements are combined in molecules that formed the building blocks of the planets.

Table 1.2 The approximate atomic composition of a human being compared with astronomical objects. The number preceding each element in the left-hand column is the number of protons in the nucleus.

	Abundance (% by mass)*			
Element	**Sun (photosphere)**	**Whole Earth**	**Earth's crust**	**Human**
1 Hydrogen	**74**		0.1	**10**
2 Helium	**25**			
8 Oxygen	0.6	**30**	**47**	**61**
6 Carbon	0.2	0.1	0.1	**23**
26 Iron	0.1	**32**	5.1	
10 Neon	0.1			
12 Magnesium	0.1	**15**	2.1	
7 Nitrogen	0.1			2.6
14 Silicon	0.1	**16**	**28**	
16 Sulfur		0.6	0.1	0.2
28 Nickel		1.8		
20 Calcium		1.7	3.7	1.4
13 Aluminium		1.6	**8.1**	
11 Sodium		0.2	2.8	0.1
24 Chromium		0.5		
15 Phosphorus		0.1	0.1	1.1
25 Manganese		0.2	0.1	
17 Chlorine			0.1	0.1
27 Cobalt		0.1		
19 Potassium			2.6	0.2
22 Titanium		0.1	0.6	

* Abundances of less than 0.1% are not shown. The biggest contributors are in **bold**. Totals do not add up to 100% because of rounding and missing elements.

The Earth is a rocky body, so elements that form minerals and rocks (oxygen, silicon and metals such as iron and magnesium) are the most common. Helium is an inert gas (as are neon and argon), which means that it does not react with other atoms to form molecules, and therefore does not contribute significantly to the rocky material of the Earth. The gas giant planet Jupiter has a composition much closer to that of the Sun; its vast atmosphere is composed of mainly hydrogen and helium. The reasons for these differences in composition of the planets will be presented in Chapter 7.

The elemental composition of the human body is dominated by hydrogen, carbon and oxygen, three of the four most abundant atoms in the Sun Water is formed from hydrogen and oxygen, and carbon is the key to forming highly complex molecules. These organic (carbon containing) molecules provide the framework for constructing the complex structures present in the human body and for carrying the information that allows humans to grow and reproduce (one of the definitions of life).

■ Helium is the second most abundant element in the Universe but is not a significant component of the human body. Why is this?

□ Because helium is an inert gas it does not form into molecules that are present in the human body.

In Chapter 8 you will investigate the requirements for these complex molecules that are essential for life (as we know it) to exist. Understanding how those conditions are satisfied on the Earth is the first step in attempting to find environments elsewhere where life may exist.

This book and other resources in the module provide a very brief introduction to the subject. Even so, it will take you not only on a journey through space but also on a trip through time. Current views of different parts of the Universe offer a wonderful spectacle, but all astronomers know that, as they peer across vast cosmic spaces, they also look back over great reaches of cosmic time. This is an unavoidable consequence of the finite **speed of light**. The light seen today from the most distant observable galaxies was emitted over 12 billion years ago. The earliest signals of any kind that can be detected (a particular kind of microwave radiation that is the remnant of the **Big Bang** that is believed to have formed the Universe) originated over 13 billion years ago. The Sun and Earth were formed around 4.6 billion years ago, and life has developed on Earth within the last 3 billion years Only in the last million years of this vast timescale, did humans evolve on Earth, and only in the last century have they been able, in theory, to communicate their presence to other possible inhabitants of our galaxy. One of the most exciting recent developments in astronomy has been the detection of planetary systems around other stars. Many astronomers believe firm evidence for the presence of extraterrestrial life, either on another planet in our Solar System or orbiting a different star, will be found during our lifetimes. The prospects for finding extraterrestrial intelligent life appear to be much more remote. In this module you will discover how all stages in the evolution of the Universe (Figure 1.17), from the formation of atoms to their incorporation in galaxies, stars and planets, have played a role in determining the environments where life could exist.

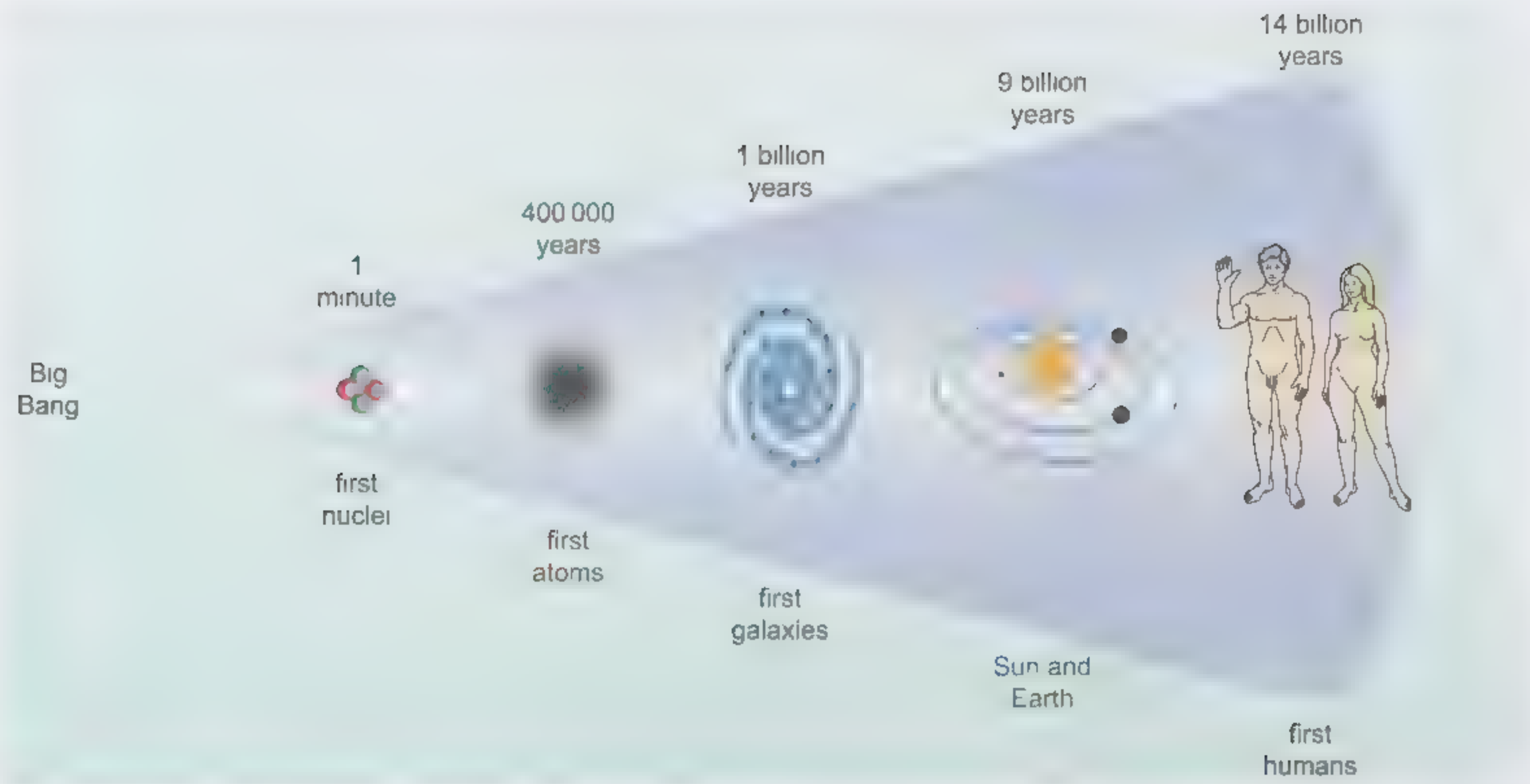


Figure 1.17 Cosmic time, from the Big Bang, about 13.7 billion years ago, to the present. The first galaxies may have formed less than one billion years after the Big Bang. The Sun and the Earth formed about 4.6 billion years ago, when the Universe was about 9 billion years old.

End-of-chapter questions

The answers to end-of-chapter questions can be found at the back of this book.

Question 1.1 What would be the diameter of the Sun if it was represented in the model of the Solar System described in Table 1.1?

Question 1.2 Distances to nearby stars are a few light-years. Calculate how long it takes light to reach the Earth from the Sun and hence explain why the light-year is not a useful unit for measuring the distance to the very nearest star, the Sun.

Question 1.3 Two stars, X and Y, have the same size and colour but X is four times further away from an observer on Earth than Y. How will the apparent brightness of the stars compare to the observer?

Question 1.4 The majority of the Earth's rocky surface is comprised of silicate rocks (compounds containing silicon and oxygen together with metals such as aluminium, calcium, magnesium and iron). The core of the Earth is believed to be predominantly composed of iron. What evidence is there in Table 1.2 to support this?

At the end of each chapter you will be directed to the module website, where there are additional activities. You should do this now for Chapter 1. You may like to look at these activities at the start of each chapter to get an idea of what they involve.

Activity 1.2 - 30 mins
Activity 1.3 - 25 mins
Activity 1.4 - 10 mins

Chapter 2 Measuring the Universe

2.1 Introduction

Chapter 1 summarised some of the properties of the Universe. How were they derived?

Determining distance is the key to deriving many of the other properties of objects in the Universe. By observing everyday objects it's possible to judge distances of a few metres by using stereo vision. We use visual clues to determine size or distance on larger scales (e.g. knowing how tall a typical person is makes it possible to judge their distance from their apparent size). In astronomy, distances are generally so great that we have to use similar clues to determine properties. A star is seen as a point of light – is it a distant, intrinsically luminous star, or is it a nearby intrinsically faint star? Without other information (or assumptions) we cannot say which.

One of the key assumptions astronomers make is that, in the absence of evidence to suggest otherwise, the properties of certain objects are the same, wherever they may be found in the Universe. Knowing the typical size of a tree makes it possible to use the apparent size of a tree on a distant hill to estimate the distance to that hill. In the same way, knowing the typical size of a certain type of galaxy makes it possible to use the apparent size of such a galaxy in a cluster of galaxies to estimate the distance to that cluster. This is only an 'estimate' because we know that any individual tree (or type of galaxy) may be smaller or larger than the typical size. This introduces an uncertainty into the distance determination.

Specifying uncertainties is very important in astronomy, as in all science and everyday life. For example, if you use a tree on a hill to estimate the distance to the hill as 5 km (see Section 2.3.2 for more details) you may decide that you'd enjoy a leisurely walk to the hill within 2 hours. However, if the tree turned out to be twice the height of a typical tree, its distance would actually be 10 km and you'd discover that the walk was much longer (or you may have decided not to walk if you'd known!). This problem could have been avoided if there were many trees on the hill, so you could estimate the distance much better by using a tree of average height. Alternatively, you could search for an object that doesn't have such a large range of heights to judge the distance.

- What properties are required for an object to be ideal for distance determination?

  The object must have a property that is constant and easily measured anywhere in the Universe, and must be present in all the locations whose distance we wish to measure.

In practice, such objects are impossible to find. Determination of distances across the vast scales of the Universe requires a range of different techniques. You will learn about some of these techniques in this chapter, and others as we investigate the properties of stars and galaxies in later chapters.

As you have seen in Chapter 1, the scale of the Universe is so vast that we need to introduce new units for measurements of size and distance. However, even with these units, making calculations can be difficult, as the example below shows.

■ The Sun is the nearest star to the Earth. How much further away is the next closest star, which lies at a distance of 1.3 pc?

The distance to the Sun is 1 AU which is 150 000 000 km. One parsec is about 3.26 ly and 1 ly is 9467 280 000 000 000 metres (Section 1.2), so the nearest star is (1.3 × 3.26 × 9467 280 000 000 000 m)/(150 000 000 × 1000 m) times further away. This is approximately 270 000 times further away.

If you tried to do the calculation above using an electronic calculator, you may have encountered difficulties with entering some of the very large numbers. To avoid such problems **scientific notation** is used.

Table 2.1 shows how it works: for example, 100 = 10 × 10, which can also be written as 10^2 and read as 'ten to the power of two'; 1000 = 10 × 10 × 10, or 10^3; and so on. In scientific notation, five thousand can be written as 5×10^3 (five times one thousand) and five hundred million as 5×10^8, which is much more compact, and easier to read than 500 000 000 (once you are used to it). You don't have to count the zeros because the 'power' (the small number, called the **exponent**) tells you how many there are.

Table 2.1 Powers of ten.

100 =	10 × 10 =	10^2
1000 =	10 × 10 × 10 =	10^3
10 000 =	10 × 10 × 10 × 10 =	10^4
100 000 =	10 × 10 × 10 × 10 × 10 =	10^5
1000 000 =	10 × 10 × 10 × 10 × 10 × 10 =	10^6
and so on ...		

In scientific notation, 1 astronomical unit could be written as 1.5×10^8 km, or in metres as 1.5×10^{11} m. The convention is to write just one digit before the decimal point – writing 15×10^{10} m would not be incorrect, just unconventional.

■ Using scientific notation, write down the distance of 1 ly in metres.

1 ly is 9467 280 000 000 000 m, which is $9.467\,28 \times 10^{15}$ m.

Your calculator may be different.

To enter such a number on a calculator, you need to type in 9.46728 then press the exponent button (EE or EXP), then 15.

Currently, the word 'billion' is used to mean 10^9 (or 1000 000 000), which is one thousand million.

As you saw in Chapter 1, very small numbers are needed to describe the sizes of atoms. As with writing very large numbers, scientific notation can be used to make the numbers more compact. Table 2.2 shows how very small numbers can be written by extending the pattern from Table 2 1. You may be surprised to see that $10^0 = 1$, but that follows inevitably if the pattern is continued downwards, dividing by 10 each time. For small numbers, note that the power is the same as the number of zeros (including the one before the decimal point) before the first non-zero digit (*not* the number of zeros after the point) For example, 0.0001 is 10^{-4}.

Table 2.2 Powers of ten including small numbers.

.. continues for smaller numbers		
0.0001 =	1/10 000 = 1/(10 × 10 × 10 × 10) =	10^{-4}
0.001 =	1/1000 = 1/(10 × 10 × 10) =	10^{-3}
0.01 =	1/100 = 1/(10 × 10) =	10^{-2}
0.1 =	1/10 =	10^{-1}
1 =	1 =	10^0
10 =	10 =	10^1
100 =	10 × 10 =	10^2
1000 =	10 × 10 × 10 =	10^3
10 000 =	10 × 10 × 10 × 10 =	10^4
100 000 =	10 × 10 × 10 × 10 × 10 =	10^5
1000 000 =	10 × 10 × 10 × 10 × 10 × 10 =	10^6
continues for larger numbers		

Small numbers are written in scientific notation in just the same way as large numbers, i.e. with one figure before the decimal point. For example, 0.0003 is written 3×10^{-4}, and 0.0076 is written 7.6×10^{-3}. To enter such numbers on a calculator, you need to use the 'change sign' button (labelled + −) *not* the minus button; so, to enter 7.6×10^{-3} you would type in 7.6 then press the exponent button (EE or EXP), then + and then 3. Note that this may vary depending on your calculator.

■ The radius of a proton is 877 attometre (am). Write this size in metres in scientific notation.

An attometre is 1 m/1000 000 000 000 000 000 (see Section 1.2). So 1 am is 0.000 000 000 000 000 001 m = 10^{-18} m. The radius of a proton is therefore 877×10^{-18} m, which is 8.77×10^{-16} m.

The SI unit of length, and those derived from it, described in Section 1.2 can therefore be defined much more succinctly in scientific notation as shown in Table 2.3.

Table 2.3 Units of distance expressed in scientific notation in ascending order. Those in bold are in common use. Note that 1 micrometre (μm) is often referred to as 1 micron.

SI unit	Shortened name	Value in metres in scientific notation	Other units commonly used in astronomy
1 attometre	1 am	10^{-18} m	
1 femtometre	1 fm	10^{-15} m	
1 picometre	1 pm	10^{-12} m	
			1 **ångström** = 1 Å = 10^{-10} m
1 **nanometre**	1 nm	10^{-9} m	
1 **micrometre**	1 μm	10^{-6} m	
1 **millimetre**	1 mm	10^{-3} m	
1 **metre**	1 m	10^{0} m	
1 **kilometre**	1 km	10^{3} m	
1 megametre	1 Mm	10^{6} m	
1 gigametre	1 Gm	10^{9} m	
			1 **astronomical unit** = 1 AU = 1.5×10^{11} m
1 terametre	1 Tm	10^{12} m	
1 petametre	1 Pm	10^{15} m	
			1 **light-year** = 1 ly = 9.47×10^{15} m
			1 **parsec** = 1 pc = 3.08×10^{16} m
1 exametre	1 Em	10^{18} m	
			1 **kiloparsec** = 1 kpc = 3.08×10^{19} m
1 zetametre	1 Zm	10^{21} m	
			1 **megaparsec** = 1 Mpc = 3.08×10^{22} m
1 yotametre	1 Ym	10^{24} m	

Distances across our galaxy are more often referred to in kiloparsecs (i.e. 10^3 pc) and distances to other galaxies in megaparsecs (i.e. 10^6 pc).

More about units

Derived units, such as those for speed that result from combining base units, are written in a particular way in the SI system. The box on page 15 introduced you to powers of numbers; here you will extend this to powers of units. Speed is distance divided by time. So the SI unit of speed is metre per second or m/s. In the SI system this unit would be written as m s^{-1}. This is because metre/second is the same as metre × (1/second) which can also be written metre × (second to the power −1) (or m s^{-1} for short).

The area of a rectangle is length × width. The unit of area is therefore the unit of length × the unit of width, e.g. metre × metre, or square metre, written m^2. The density of a body is its mass divided by its volume. What is the unit of density? Density = mass/volume. The SI unit of mass is the kg. The unit of volume is the cubic metre (e.g. length × width × height, for a block), written m^3. The unit of density is therefore kg/m^3, written kg m^{-3}.

Before looking at how astronomers determine the distances to galaxies and stars as well as other properties such as temperature and composition from observation, you'll first need to understand the properties of light (or more correctly *electromagnetic radiation*) that are fundamental to all observations.

2.2 The nature of light

2.2.1 Light – waves and particles

Most astronomical observations are made by the detection of **radiation** used here to mean that which is radiated from a source and travels through space. It is often used as an abbreviation for **electromagnetic radiation**. Human eyes are sensitive to light that comprises the familiar rainbow of colours but these colours are just a tiny part of the **electromagnetic spectrum**, which includes X-rays, microwaves and radio waves (see Figure 2.1).

One way to describe the different components of the electromagnetic spectrum is in terms of waves. A **wave** may be defined as a *periodic* (regularly repeating) disturbance that transports *energy* from one place to another. For instance, a stone dropped into the centre of a pond generates waves on the water surface, which travel outwards and eventually cause a cork at the edge of the pond to bob up and down with a regular motion. Similarly, a sudden motion of part of the Earth's crust generates seismic waves that travel through the Earth, and may cause damage to buildings some distance away on the surface. Another image that the word 'wave' often conjures up is that of water waves on the sea.

If you've ever been on a beach you will have seen or heard waves break onto the shore with a fairly regular time interval between each 'crash' and the next. Each crash represents one wave crest breaking onto the shore, and the time

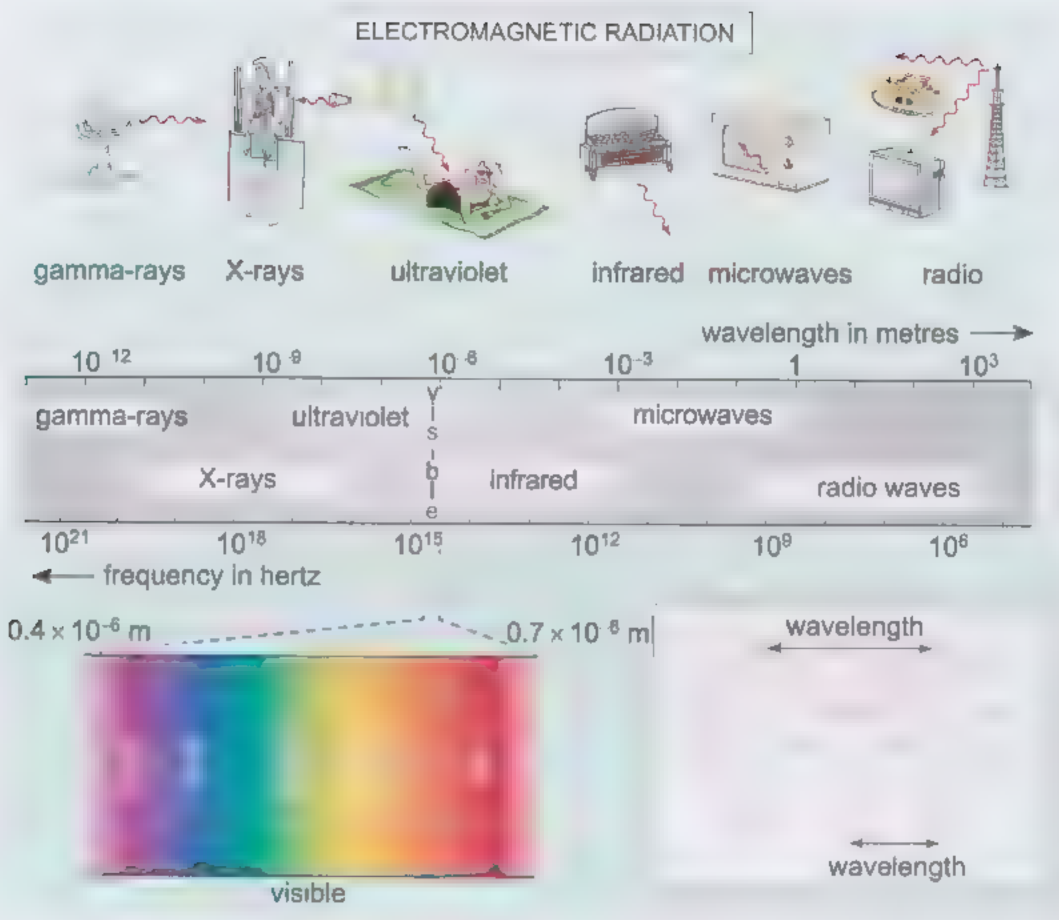


Figure 2.1 The electromagnetic spectrum. The various kinds of electromagnetic radiation can be distinguished by their *wavelength* or *frequency*. In the case of visible light, different wavelengths are perceived as different colours but there is no fundamental distinction between the parts of the spectrum.

Increasing frequency = decreasing wavelength

wavelength = $\frac{\text{speedoflight}}{\text{frequency}}$

interval between two of them is known as the period of the wave. In general, the period of a wave may be defined as the time between one part of the wave profile (say the crest) passing a fixed point in space and the *next* identical part of the wave profile (the next crest) passing the same fixed point. In the example here, the fixed point is the shoreline.

As well as being periodic with time, a wave is also periodic as far as its spatial extent is concerned. The word 'wave' is often used to describe a *single* crash onto the beach, but it really refers to the entire sequence of crests and troughs, stretching away into the distance. The distance between one wave crest and the next is known as the **wavelength** of the wave. In general, the wavelength of a wave is defined as the distance between one part of the wave profile, at a particular instant in time, and the *next* identical part of the wave profile at the same instant of time. Two adjacent crests of the wave are a convenient pair of locations to use for this definition, although any pair of similar points will do. This is shown in the lower right-hand panel of Figure 2.1.

You can think of electromagnetic waves as similar to waves on the ocean. In the case of ocean waves, the wave motion takes place on the surface of the water, whereas in the case of electromagnetic waves the wave motion takes the form of varying electric and magnetic disturbances. The exact nature of these electromagnetic waves is not important for this discussion, and if you keep in mind a visual image of water waves, that will be fine.

The key thing is that each colour of light corresponds to electromagnetic waves of a different wavelength. Visible light spans the range of wavelengths from about 400 nanometres, (i.e. 4×10^{7} m) to 700 nanometres or 7×10^{-7} m. Violet light has the shortest wavelength and red the longest (Figure 2.1). At even longer wavelengths are found first infrared radiation, then microwaves and radio waves, whereas at wavelengths shorter than the visible are found ultraviolet radiation, then X-rays and gamma-rays.

Radiation can also be characterised by its **frequency**. Using the analogy of waves arriving at the beach, the frequency is the number of waves reaching the beach per second. If the waves are travelling at a speed, v (measured in metres per second) and the wavelength of the waves is denoted by the Greek letter lambda, λ, (measured in metres), then the frequency, f, is equal to v divided by λ The frequency is therefore measured in units of 1 second called hertz (Hz). As electromagnetic radiation travels at the speed of light, often represented by the letter c, its frequency f is therefore defined by:

Scientists often use letters or symbols to represent quantities. The Greek letter lambda is pronounced 'lamb-der'.

frequency = speed of light/wavelength or $f = c/\lambda$.

- What is the frequency range of visible light?

  Visible light has a wavelength range of 4×10^{-7} m (violet) to 7×10^{-7} m (red). The frequency of violet light is given by

  $f = (3 \times 10^{8}\ \text{m s}^{-1})/(4 \times 10^{-7}\ \text{m}) = 7.5 \times 10^{14}$ Hz.

  The frequency of red light is given by

  $f = (3 \times 10^{8}\ \text{m s}^{-1})/(7 \times 10^{-7}\ \text{m}) = 4.3 \times 10^{14}$ Hz.

Dividing a number in m s^{-1} by a number in m leaves you with a number with units s^{-1}, also known as Hz (hertz).

Although electromagnetic radiation travels from place to place like a wave, it is emitted or absorbed by matter as if it is composed of a stream of particles, called **photons**. There is no conflict between these two approaches: it's just that one picture (waves) is useful for describing the way electromagnetic radiation propagates and the other picture (particles) is useful for describing the way it interacts with matter. This double description is referred to as **wave–particle duality**. We therefore refer to X-ray photons, visible light photons or microwave photons as the particles corresponding to electromagnetic radiation at each of these wavelengths. Like any particles, photons can be characterised by how much energy they carry. Photons of shorter wavelength (higher frequency), such as X-rays and gamma-rays carry more energy than visible light, whereas photons with longer wavelengths (lower frequencies) carry less energy than visible light photons. The energy of a photon is directly proportional to the frequency so it can be calculated using the equation:

$\varepsilon = h\,f$ is the same as saying 'ε equals h multiplied by f'.

energy of photon = $\varepsilon = h\,f$

where h is a constant called *Planck's constant* and has a value of $6.63 \times 10^{\ 34}$ J s. Since we know the relationship between frequency and wavelength, the energy of a photon can also be expressed as

energy of photon = $\varepsilon = hc/\lambda$

■ What is the energy range of visible light photons?

The result of the calculation above was that visible light has a frequency range from 7.5×10^{14} Hz for violet light to 4.3×10^{14} Hz for red light. The corresponding photon energies are therefore

$\varepsilon = hf = (6.63 \times 10^{-34}\ \text{J s}) \times (7.5 \times 10^{14}\ \text{Hz}) = 5.0 \times 10^{-19}$ J for violet light and

$\varepsilon = hf = (6.63 \times 10^{-34}\ \text{J s}) \times (4.3 \times 10^{14}\ \text{Hz}) = 2.9 \times 10^{-19}$ J for red light.

Multiplying a number in J s by a number in s^{-1} (Hz) leaves you with a number with units J.

Radiation from a bright body, such as a planet, star or galaxy, naturally consists of a mixture of many wavelengths, often covering an almost continuous range. When referring to the **spectrum** of a particular body it usually means the range of wavelengths coming from that body, along with additional information about the 'amount' or 'intensity' of the radiation in any narrow band of wavelengths (a measure of the energy carried by that band of wavelengths).

The spectrum of visible light from a celestial body can be examined by passing the light from that body through a narrow slit and then allowing it to pass through a triangular glass prism (see Figure 2.2). The prism *disperses* the light, causing slightly different wavelengths to travel in slightly different directions. As a result, the light spreads out to form a rainbow-like band ranging from red to violet. The variation of intensity with wavelength is revealed by the relative brightness of the different parts of the spectrum.

The visible spectrum of a hot, dense body consists of a continuous band of colours (Figure 2 2a), which we call a **continuous spectrum**. If, however, a cloud of gas is interposed between the source and the observer, the atoms in that cloud will absorb certain characteristic wavelengths, producing a corresponding pattern of relatively dark 'absorption lines' that cross the spectrum, destroying its smooth continuity (Figure 2 2b). This is called an **absorption spectrum**. The energy carried by the 'missing' wavelengths is absorbed by the cloud, but only temporarily. The cloud will soon lose that energy, often by emitting just the same wavelengths of radiation that it absorbed earlier. However, the emitted radiation generally goes in all directions, so it may be seen by observers who are not looking towards the hot, dense body (Figure 2.2c). The spectrum seen by such observers consists of relatively bright 'emission lines' against a generally dark background and is called an **emission spectrum**. These characteristic wavelengths can be used to determine the composition of the gas cloud (see Section 2.3).

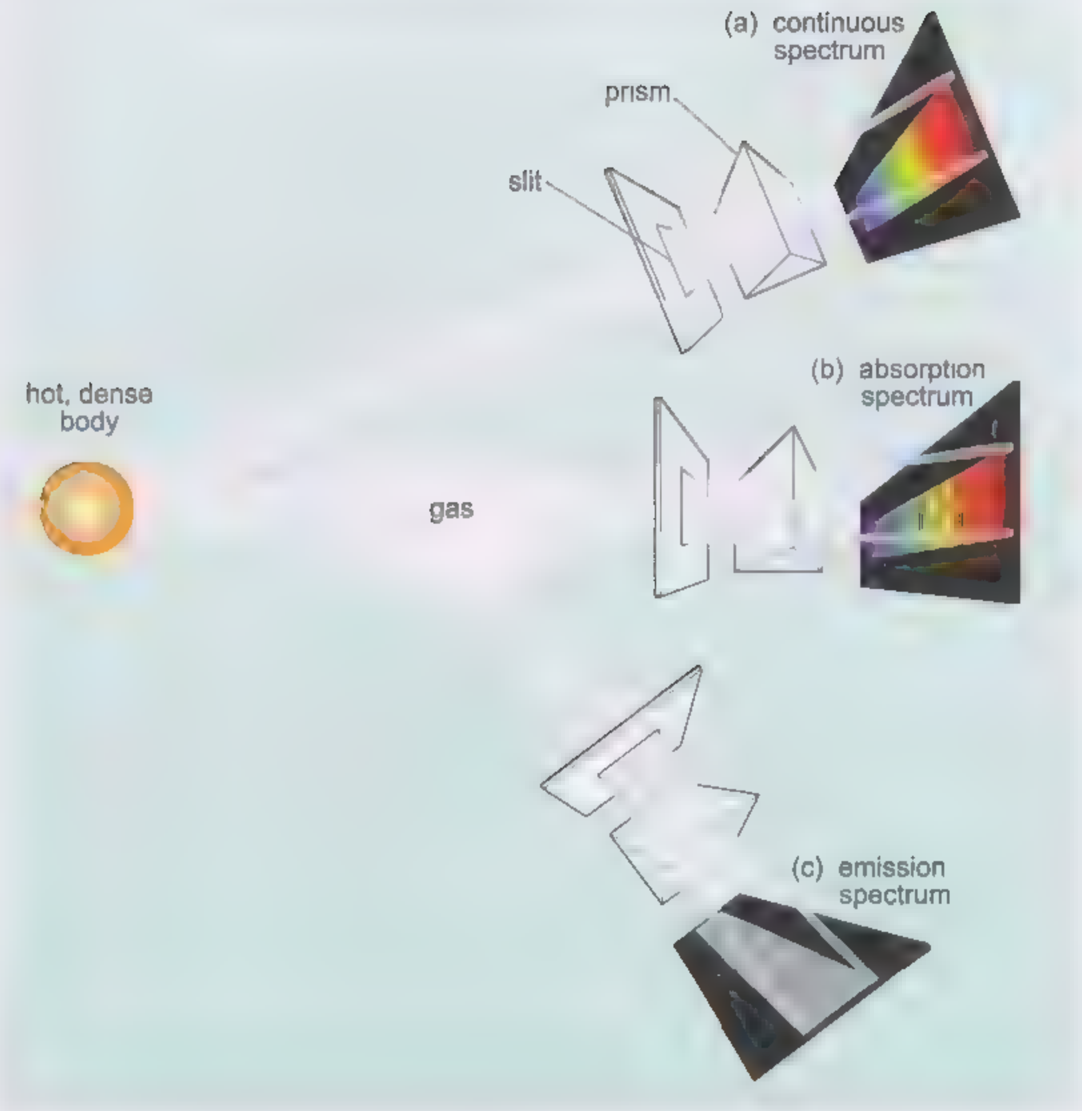


Figure 2.2 Three kinds of spectra. (a) The unbroken, continuous spectrum of a hot, dense body. (b) The absorption spectrum when light with a continuous spectrum passes through a cloud of gas. (c) The emission spectrum of light radiated from a cloud of gas.

Stars typically produce absorption spectra. A continuous spectrum is produced by the dense gas in the photosphere of the star and absorption lines result from the cooler atmosphere lying above it.

Spectra are also used to study the movements of astronomical bodies (Section 2.5.3) and even to deduce details about their temperatures (Section 2.3) and physical properties. Spectra are estimated to account for more than 90% of all available astronomical information; this estimate emphasises the intimate linkage between matter and radiation. Thanks to space telescopes and sophisticated radiation detectors, all parts of the electromagnetic spectrum can now be used for astronomical observations, and each kind of radiation has contributed to the growth of astronomical knowledge.

2.2.2 The Universe at different wavelengths

As you've just learnt, astronomers use telescopes that are sensitive to different regions of the electromagnetic spectrum in order to study the Universe. But there is a problem here – human eyes aren't sensitive to these wavelengths, so how can anyone view the data collected by these telescopes? The answer to this is that astronomers can make *false colour* images. All this means is that astronomers represent an image of an object that was observed at a particular wavelength of light for which the human eye is not sensitive, using a colour of light at which the human eye *is* sensitive. An everyday example of this is seen in the thermal imaging cameras that the emergency services use to search for people in the dark via the heat from their bodies – these convert infrared light (which our eyes can't see) into visible light.

Let's take this approach further. Televisions produce full colour images by combining images taken in red, green and blue light. In a similar way, astronomers can produce a false colour image from a combination of observations taken at different wavelengths. Each of these wavelengths is represented by a visible light colour and the resultant monochrome (or single colour) images are combined to form a false colour image. Typically, astronomers will code the differing wavelengths of light they observe in such a way that the shortest wavelength observed is assigned the shortest wavelength of visible light (violet or blue) in the false colour image, and the longest wavelength observed is assigned the longest wavelength of visible light (red) to make the image comparable to how we observe the everyday world. A beautiful example of this is the image of the Whirlpool galaxy shown in Figure 2.3.

Why would anyone want to look at the Universe at different wavelengths? The simple answer is that astronomical objects such as planets, stars and galaxies can look strikingly different when observed in different regions of the electromagnetic spectrum. And it turns out that to fully understand such objects we need the clues provided from images taken at all these wavelengths. To demonstrate this, let's look at the Whirlpool galaxy shown in Figure 2.3. This is a famous example of a spiral galaxy – thought to be similar to our own Milky Way. Seen face on, the nested spiral arms are clearly visible in the *optical image*, delineated by the light of the millions of individual stars that are found within them. You may also be able to see two other notable features. Thin dark lanes run along the spiral arms – these are due to cold dust and gas, which absorb the light from stars behind them. There's also another, fuzzy egg-shaped object at the top of the picture, which is a smaller galaxy that is interacting with the Whirlpool. But is it possible to say anything more about this, or indeed what the Whirlpool is made from (apart from stars)? This needs observations at other wavelengths. If radiation is observed at wavelengths just longer than those at which our eyes are sensitive, it's often referred to as infrared light. As you'll learn later in this chapter, light of this wavelength is typically emitted from relatively 'cool' astronomical objects – although this is of course relative, since in astronomical terms 'cool' means temperatures of typically 'only' a few thousand kelvin or less! At such wavelengths it's possible to detect bright emission from the dust that was previously only detectable by the light it

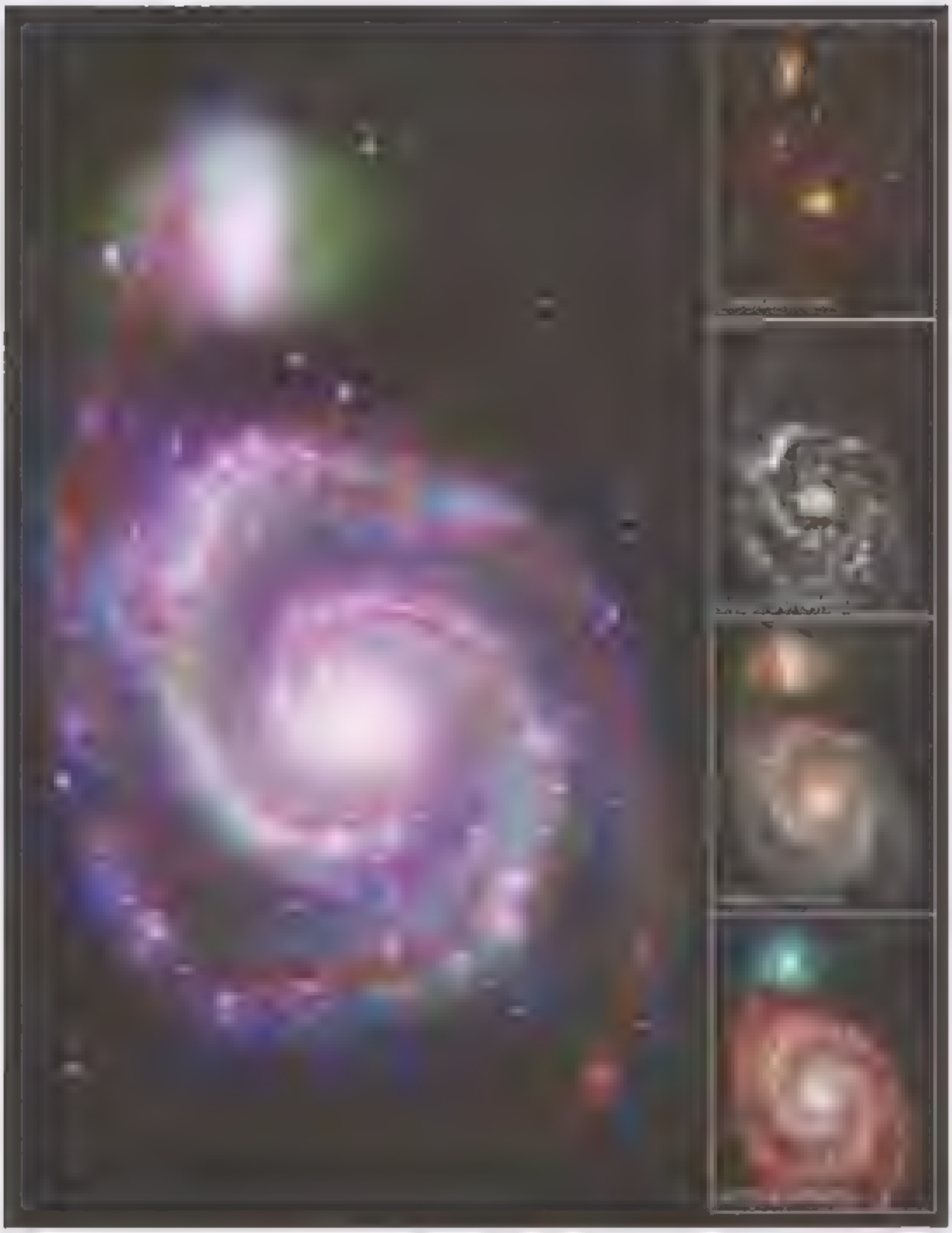


Figure 2.3 Multiwavelength observations of the Whirlpool galaxy and a smaller, companion galaxy. Main panel: A false colour image of the Whirlpool galaxy composed of X-ray (coded in violet), ultraviolet (blue), optical (green) and infrared (red) observations. The individual images that were combined to make this main image are shown in the inset panels – these also were made from multiple observations at different wavelengths and so are also false colour images. Note that 'optical' refers to light that is normally detected with an optical telescope, i.e. the visible spectrum plus parts of the ultraviolet (that is transmitted by the Earth's atmosphere) and infrared wavelengths either side.

absorbed at optical wavelengths. The dust in Figure 2.3 also traces the location of the spiral arms. At these wavelengths the stars lying in the spiral arms that were bright at optical wavelengths are hardly visible, while the ellipsoidal galaxy at the top of the image is still clearly visible, even if individual stars may not be detected due to its distance.

Observing these galaxies at shorter (ultraviolet) wavelengths tells a different tale. Here you can clearly see emission from stars along the arms of the Whirlpool galaxy, but the smaller galaxy is now completely invisible! These differences tell us that the stars found in each galaxy emit different wavelengths of light, with the stars in the Whirlpool galaxy predominantly radiating light at shorter wavelengths. Later in this chapter, you'll learn more about what this can reveal about the nature of the stars in both galaxies.

Finally, one can look at the galaxies at X-ray wavelengths, finding that they look completely different. Instead of the beautiful spiral structure there are only a few diffuse blobs and isolated sources. Clearly, whatever is emitting such high-energy radiation is not directly related to the stars that are visible at other wavelengths.

You can now see why astronomers make use of observations of multiple wavelengths. Simply observing the Whirlpool galaxy and its companion at optical wavelengths wouldn't have revealed that the stars in each galaxy look very different from one another, nor would it have revealed the X-ray emitting objects that are present but aren't visible at other wavelengths.

2.2.3 Observing at different wavelengths

How do astronomers accomplish this? You're probably familiar with instruments designed to look at optical wavelengths – these are simply traditional telescopes. Likewise, telescopes built to look at radio wavelengths are immediately recognisable with their iconic dish design (Figure 2.4d). The one thing that both these have in common is that they're located on the surface of the Earth and make their observations through the atmosphere. However, observatories designed to observe at other wavelengths are found in space. Given the cost and technical difficulty of building and launching such satellites, why is this the case?

The reason for this is that the atmosphere is transparent to light at some wavelengths, *but not all*. Visible light passes through the atmosphere – if it didn't, our eyes wouldn't have evolved to be sensitive to it. The atmosphere is also transparent to radio wavelengths, but light at other wavelengths is completely absorbed by the atmosphere before it reaches the surface of the Earth. The example that you are probably most familiar with is the absorption of harmful ultraviolet radiation by the ozone layer, but the atmosphere is also opaque to X-rays and light with wavelengths of one-tenth of a millimetre – in both cases, telescopes sensitive to such light must be located above the Earth's atmosphere.

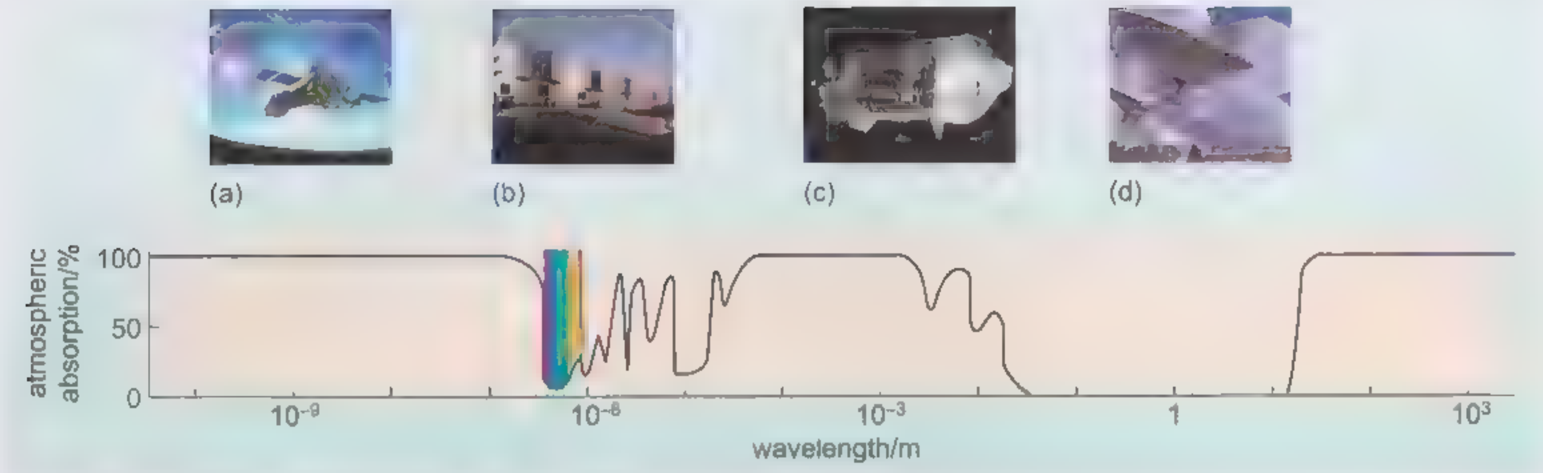


Figure 2.4 Schematic illustrating the absorption by the Earth's atmosphere at sea-level of light of different wavelengths Note that telescopes intended to observe at different wavelengths may have very different designs including (a) XMM (X-ray), (b) the Very Large Telescope (optical), (c) Herschel (sub-mm wavelengths) and (d) the Very Large Array (VLA), which is a radio telescope.

2.2.4 The invisible Sun

Figure 2.5 shows an image of the Sun recorded using instruments that are sensitive to ultraviolet radiation rather than visible light, so the colours that you see are false.

The Sun's radio waves carry much less energy than its visible light but they can readily be detected with even a small radio telescope. Fortunately for life on Earth, the Earth's atmosphere shields us all from the Sun's potentially harmful X-rays, so these can only be studied using telescopes put into orbit above the atmosphere. A combination of ground-based and space-based instruments has enabled astronomers to observe the Sun over a wide range of wavelengths and to build up a clear picture of its various emissions.

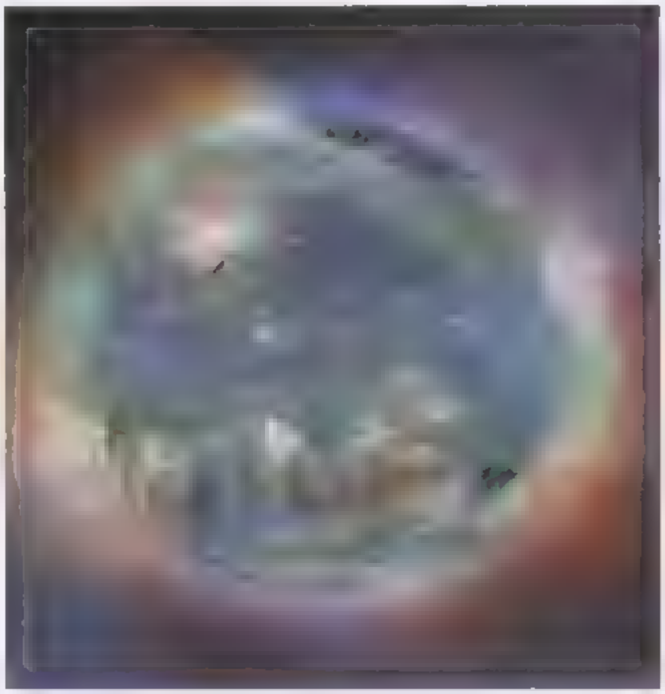


Figure 2.5 A false colour image of the Sun, taken at ultraviolet wavelengths

Activity 2.1 Examining images of the Sun

The estimated time for this activity is 20 minutes

This activity uses images from a variety of telescopes, representing various wavelengths of 'invisible' radiation (see the module website for instructions). Some of these images enhance astronomers' knowledge of particular solar features (such as prominences or sunspot groups), while others help them to observe particular regions (such as X-ray images of the corona or ultraviolet images of the chromosphere). When undertaking this activity take care not to be misled by the use of false colours.

The detailed notes for this activity can be found in the 'Activities' section of the module website.

Using the internet for updates

The Sun is constantly being watched from a variety of observatories. You can usually find recent images by searching the internet, using terms such as 'solar image' or modifications such as 'solar image, X-ray'. Some particularly useful websites are given on the module website. These sites have been chosen partly because of their reliability. By all means look for other websites but be aware that there are few guarantees of quality or reliability on the internet. Always ask yourself how much you should rely on any particular source. University and space agency websites are generally fairly reliable but even there you should exercise caution.

2.3 Measuring temperatures

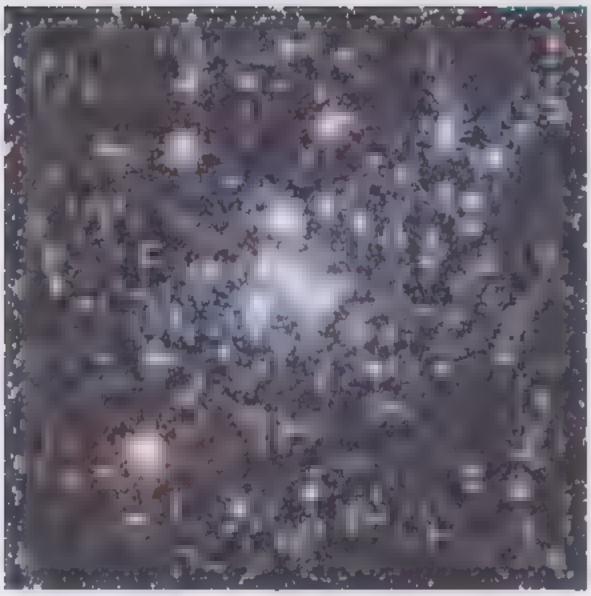


Figure 2.6 The Jewel Box (NGC4755), a young star cluster located in the constellation of Crux in the southern sky.

Earlier, this chapter touched briefly on the concept of spectra, and how they could communicate a wealth of information about the nature of astrophysical objects to us. But how do astronomers go about decoding these and what do they tell us? To investigate this, let's take a closer look at stars.

As you've already learnt, stars come in a wide range of temperatures. These differences are most apparent if we look at star clusters. These are huge collections of hundreds to millions of individual stars that formed at the same time and are held together by their combined gravitational attractions. One such cluster is the Jewel Box (Figure 2.6), composed of ~100 stars located approximately 6440 light-years away in the constellation of Crux in the Southern Hemisphere. It was discovered by Nicolas Luis de Lacaille in 1751, but named by the British astronomer John Herschel for its appearance:

> this cluster, though neither a large nor a rich one, is yet an extremely brilliant and beautiful object when viewed through an instrument of sufficient aperture to show distinctly the very different colour of its constituent stars, which give it the effect of a superb piece of jewellery

As Herschel noted, the most striking property of the stars in the Jewel Box is the range of colours, which is a direct consequence of their different temperatures. The easiest way to visualise it is to imagine that you are watching a blacksmith making a horseshoe. As he or she heats it in order to soften and hammer it into shape, it initially radiates no light at all, appearing dark against the glare of the fire. However, as the horseshoe becomes hotter it begins to glow, first a dull red and then a brighter orange and yellow before appearing an incandescent white. In fact the change in colour is a result of the increasing temperature of the iron – there is a direct relationship between the temperature of any object (such as a horseshoe, the element in an electric fire and even a star) and the colour (or wavelength) of light it emits.

The connection between temperature and electromagnetic radiation could be summed up as 'the hotter the brighter, the hotter the whiter'. In fact, this snappy summary is rather oversimplified because very hot objects give out more blue light than any other colour, so they appear blue rather than white. Also *very* hot objects give out much ultraviolet radiation, and may even emit X-rays, as well as visible light that is predominantly blue. At the other extreme, even cool objects give out some electromagnetic radiation. However, it nearly all consists of infrared and microwaves, which are invisible to human eyes, but detectable by suitable instruments, as previously mentioned. Nevertheless, in principle astronomers can use the colours of stars to estimate their temperatures.

A more exact summary is to say that *all* objects give out electromagnetic radiation; the hotter an object is, the more electromagnetic radiation it gives out altogether, and the more it gives out particularly at short wavelengths. As well as using words, this can be shown on a graph such as Figure 2.7. The Sun is a yellowish–white star, with a surface temperature of about 5800 K. The Sun is a middling sort of star. Some stars are blue–white, with temperatures of perhaps 15 000 K or more, while others are orange–red and relatively cool at only about 3000 to 4000 K. Stars of both temperature extremes are visible within the Jewel Box (Figure 2.6).

■ Can you see why the Whirlpool galaxy and its neighbouring satellite galaxy look rather different at ultraviolet, optical and infrared wavelengths?

On average, the stars in both galaxies have different temperatures. Those within the Whirlpool are hotter than their counterparts in the satellite galaxy and so emit most of their light in the ultraviolet and optical region of the electromagnetic spectrum. In contrast stars in the companion galaxy emit their energy in the optical and infrared regions and are not hot enough to emit ultraviolet radiation.

Sending starlight through a prism can provide even more information than just temperature. The effect of differing temperature on the spectra of stars is illustrated in Figure 2.8, which shows a number of spectra of stars with a wide range of temperatures (~40 000 to 3500 K). Firstly, looking at the continuum of the spectrum you can see that while all the stars emit light at wavelengths that we perceive to be green, the hotter stars emit more light at blue wavelengths than the cooler stars do, and conversely less red light. So, as expected the hottest stars will appear the bluest and the coolest appear the reddest, with the other stars gradually varying between these extremes – as indeed we see in the stars present in the Jewel Box cluster.

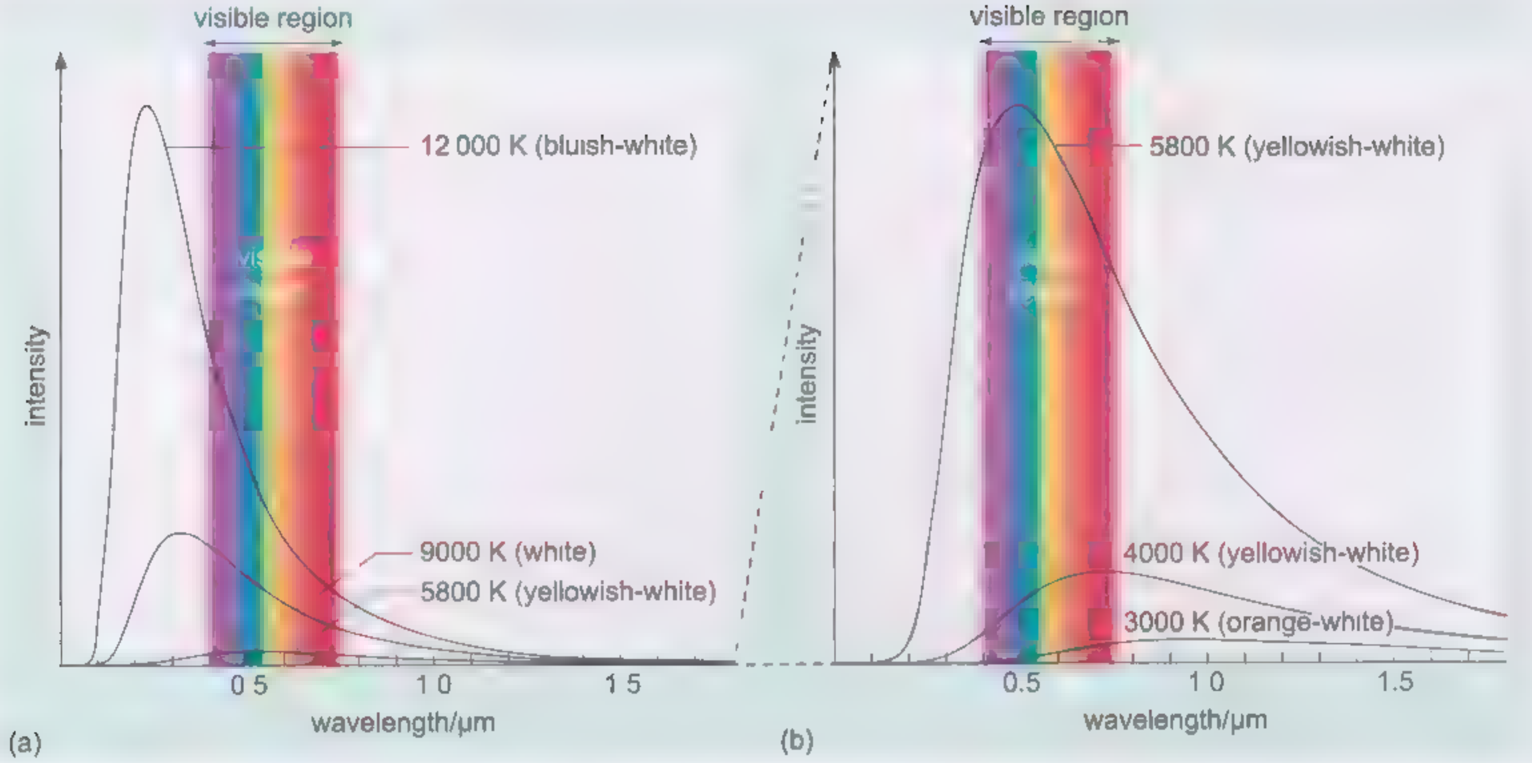


Figure 2.7 Schematic of continuous spectra of (a) hot and (b) cool objects. Each object is the same size at the same distance from a detector that measures the intensity of electromagnetic radiation. Note that the extent of the vertical axis in (a) is much greater than in (b), as shown by the dashed lines joining the two.

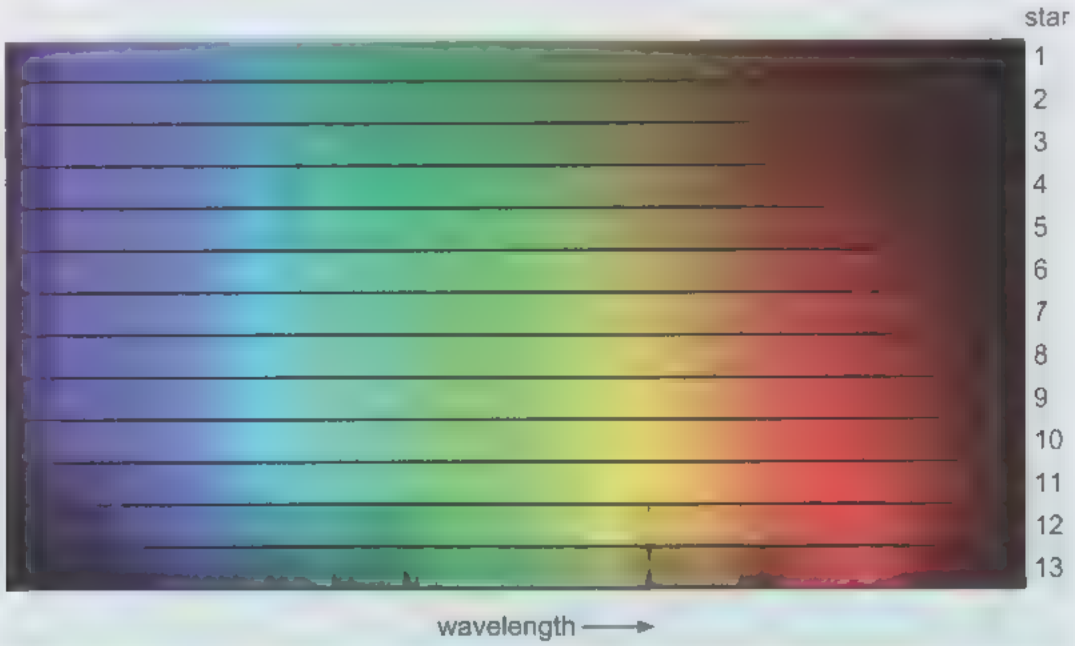


Figure 2.8 Spectra of stars of different temperatures showing prominent absorption lines. The spectra are arranged in order of decreasing temperature from top to bottom. Picture courtesy of NOAO/AURA/NSF.

However, you can also see in Figure 2.8, the presence of dark lines superimposed on the continuous spectra. The pattern of lines changes with temperature, with a general increase in frequency of occurrence for cooler stars. By studying the exact wavelengths at which these lines occur astronomers can determine both the stellar temperature and composition. Only a few lines resulting from hydrogen, helium and a few other gases are present in the spectra of the very hottest stars. However, many more lines are present in the spectra of the coolest stars. It turns out that the outer atmospheres of such stars are *so* cool that simple combinations of atoms (i.e. molecules) are present (at higher temperatures the delicate bonds holding molecules together are broken). These molecules are very efficient at absorbing light, hence leading to the large number of dark lines present. However, despite the differences in the patterns of lines in these spectra it is very important to realise that the atmospheres of all the stars have almost the same chemical composition – in this case the differences arise *solely* from the differences in temperature.

A comprehensive scheme for classifying stars on the basis of their temperature – as determined from the appearance of their spectra – was developed in the early years of the last century. In the **Harvard Spectral Classification** stars of different temperatures are assigned a letter of the alphabet. Originally, spectra were sorted into groups labelled from A (denoting the presence of strong absorption lines from hydrogen) to O and beyond (hydrogen lines weak or missing). Unfortunately, as the understanding of the physics behind the stellar spectra increased it became clear that this simple progression was inadequate to describe the nature of stars and the scheme was modified, with some groups being dropped, others merged and finally those remaining re-ordered in terms of stellar temperature. Unfortunately, this has led to a rather random collection and sequence of letters; in order of decreasing temperature, stars are classified as being of **spectral class** O, B, A, F, G, K or M (with Table 2.4 giving the average temperature for each spectral class).

Table 2.4 The spectral classes of stars and their temperatures.

Spectral class	Typical temperature/K
O	40 000
B	20 000
A	9000
F	7000
G	5500
K	4500
M	3000

Although temperature is the dominant factor, differences in the chemical composition of stars can also affect the appearance of their spectra. This can be understood by simply recognising that if, for example, hydrogen is missing from a star, then no absorption lines attributed to it will be present in the star's spectrum. So, by determining which elements are present in a star from their appearance or absence in a spectrum, astronomers can work out the chemical composition of a star (or indeed any other astronomical object). Note also that this technique is not limited to visible light, with spectroscopy used right across the electromagnetic spectrum, all the way from X-ray to radio wavelengths. It isn't quite as simple as it sounds because the spectra of stars only tell us about the parts of the star that are producing that spectrum. The observed spectra of stars therefore provide information on the atmospheres only. As you'll see later, the interiors of stars can have very different compositions.

Such observations reveal that *most* stars are composed largely of hydrogen with small amounts of other substances, although some stars in our galaxy appear to have reduced quantities of hydrogen (and helium) in their atmospheres and, conversely, have much higher percentages of heavier elements such as carbon, nitrogen and oxygen. Astronomers now believe that rather than being born with different chemical compositions these differences instead reflect changes introduced as stars use up their nuclear fuel as they near the end of their lives. The structure, composition and life cycle of stars will be covered in Chapters 5 and 6.

2.4 Measuring size with angles

2.4.1 Angles and angular size

Section 2.2 referred to observations that can only be made using sophisticated telescopes, but this section turns to an observation you can do yourself. There are two reasons for this: one is to give you experience in scientific measuring and the other is to introduce some terminology that astronomers use frequently.

The image in Figure 1.9 was taken during a total eclipse of the Sun, in which the Moon blocked out light from the Sun's photosphere, enabling the chromosphere and the corona to be seen. This happens because of a remarkable coincidence. The Sun is very much bigger than the Moon – about 400 times bigger in diameter – but it is also very much further away, by almost exactly the same factor. This means that the Sun and the Moon *appear* the same size in the sky: that is, the Sun and the Moon have the same **angular size**. Figure 2.9 illustrates this idea by showing lines drawn from an observer's eye to the extreme edges of objects at various distances. The angle between the lines determines the angular size of the objects, which would be 10° in all three cases, according to the observer as shown. Angular size thus depends on an object's actual size and its distance from the observer's eye.

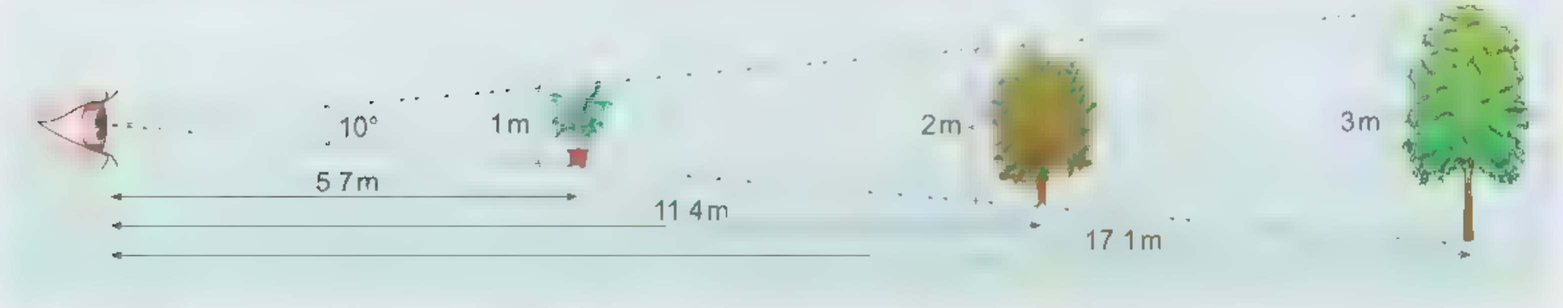


Figure 2.9 The observed angular size of an object depends on the size of the object and its distance from the observer.

You are probably used to angles measured in degrees (360° for a full circle, 90° for a right angle, and so on). However, although the objects studied by astronomers are very large, they are at such vast distances that their angular sizes are often very small. These small angles could be written as fractions of a degree or their decimal equivalents but, in practice, subdivisions of degrees are used, known as minutes of arc (or *arcmin*) and seconds of arc (or *arcsec*). A degree can be divided into 60 minutes of arc, and a minute of arc can be further divided into 60 seconds of arc. A single tick mark is used to represent arcmin, so 1/60 of a degree is written 1'. A double tick mark denotes arcsec, so 1/3600 of a degree (1/60 of an arcmin) is written 1".

Angular size is a very useful quantity in astronomy, since angular sizes are *very* much easier to measure than actual sizes because they refer simply to how large an object appears in the sky The Sun and the Moon both have an angular size of about half a degree, which is 30 arcmin or 30'.

■ How would you *define* the angular size of the Moon?

It is the angle between two imaginary lines drawn from an observer's eye to the extremities of the disc of the Moon. (More formally, it is the angle *subtended* at the eye of the observer by the *diameter* of the Moon.)

2.4.2 Angular size, actual size and distance

The angular size of an object is determined uniquely by its actual size and its distance from the observer. For an object of fixed size, the *larger* the distance, the *smaller* the angular size. For objects at a fixed distance, the *larger* the actual size of an object, the *larger* its angular size. For objects with small angular sizes, such as typical astronomical objects, the precise relationship between angular size, actual size and distance is well approximated by the equation:

$$\text{angular size} = \frac{\text{actual size}}{\text{distance}}$$

However, when using this equation you must be very careful about the units in which quantities are measured. If the actual size and the distance are measured in the same units (metres or kilometres, or anything else as long as it is used for both quantities), the angular size that you calculate will be in measured units called *radians*. One radian is equal to a little more than 57° so, in order to obtain angular sizes in degrees, the following approximation can be used (as long as the angular size is not too great):

$$\text{angular size in degrees} = 57^\circ \times \frac{\text{actual size}}{\text{distance}}$$

The next question asks you to apply this expression to Figure 2.9.

■ Calculate 57° × (actual size/distance) for each of the objects in Figure 2.9.

□ The values are 57° × (1 m/5.7 m), 57° × (2 m/11.4 m) and 57° × (3 m/17.1 m), which is 10° in each case (as expected).

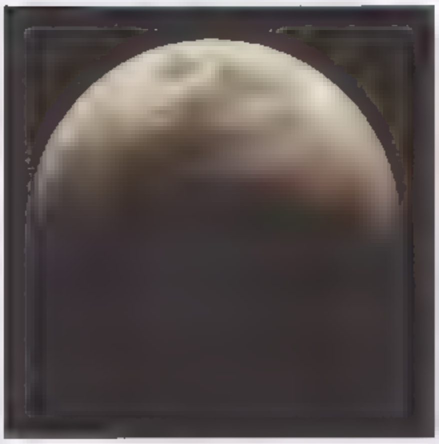


Figure 2.10 A partial eclipse of the Moon.

To do the next activity, you need to know that the Moon's diameter is 3476 km. You may wonder how this can be measured from the Earth. In principle, it is a surprisingly easy measurement to make. First, you have to find the diameter of the Earth, which can be worked out by measuring how much its surface curves. This measurement was made in about 235 BC by the Greek astronomer Eratosthenes, and his value was quite close to our modern measurement of 12 756 km (for the equatorial diameter, which is slightly bigger than the polar diameter). The sizes of the Earth and the Moon can be compared by looking at the Earth's shadow on the Moon's surface during a partial eclipse of the Moon (see Figure 2.10). A careful measurement of this kind reveals that the Earth's diameter is 3.67 times that of the Moon.

Activity 2.2 The distance to the Moon

The estimated time for this activity is 30 minutes

This activity needs to be done when the Moon is clearly visible in the sky. (It need not be done at night and, in fact, can be easier in the day or at twilight.)

DO NOT ATTEMPT THIS ACTIVITY ON THE SUN.

You will need the following items:

- a selection of round coins (e.g. UK 1p, 5p and 10p);
- a straight rod (e.g. a piece of dowelling or a garden cane) at least 2 m long;
- a tape measure at least 2 m long;
- a ruler marked in centimetres and millimetres;
- some Blu-Tack® or plasticine;
- a pocket calculator.

Set up an arrangement with a coin fixed to a rod so that the coin just 'eclipses' the Moon. Figure 2.11 shows one possible set-up.

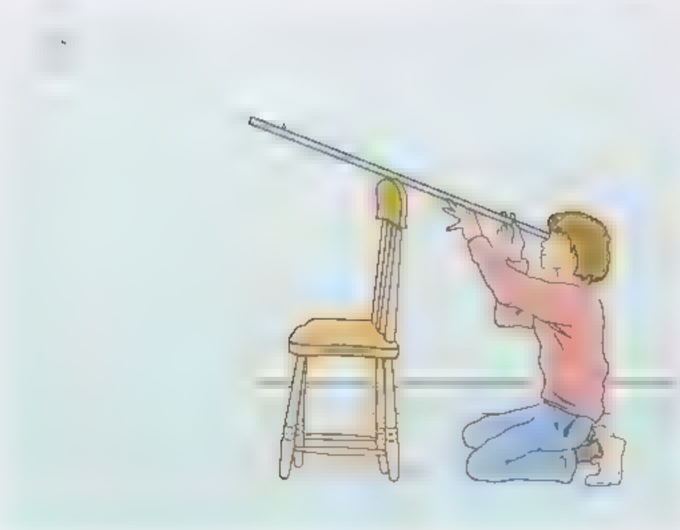


Figure 2.11 One possible arrangement for eclipsing the Moon.

Observing from one end of the rod, try different coins until you find one that is the right size to eclipse the Moon when fixed somewhere on the rod. Then adjust the position of the coin until it just blocks your view of the Moon. (This is less easy than it sounds, as there will always be some haze visible around the edge of the coin – try to get the best match.)

Measure the distance from the coin to the end of the rod where you have placed your eye, and measure the coin's diameter. Record your values here.

Diameter of coin = ... mm.

Distance of coin = ... mm.

You now have the measurements that will enable you to calculate the angular size of a coin that has the same angular size as the Moon.

Use your two measurements on the coin to calculate its angular size in degrees, using the formula introduced earlier, adapted to the current case, i.e.

angular size of coin = 57° × (diameter of coin/distance of coin) = ...

Your answer should be about half a degree (0.5°). Any value between 0.4° and 0.6° is fine. This is also your measurement of the angular size of the Moon. So, write down:

angular size of Moon in degrees = ...

The next step is to calculate the distance to the Moon. Just as for the coin:

angular size of Moon in degrees = 57° × (diameter of Moon/distance of Moon).

This expression can be rearranged to give:

distance of Moon = 57° × (diameter of Moon/angular size of Moon in degrees).

(Take this on trust if you cannot see it.) Now calculate the distance to the Moon, using your value for its angular size and 3476 km for its diameter.

Distance of Moon = 57° × (......................./...................) = ... km.

You might like to compare your result with the accurately measured value of the Moon's distance: 384 500 km. It is unlikely that you obtained exactly this value, but you probably got something in the range 300 000 to 500 000 km, which is pretty good for a quick and fairly rough measurement.

The technique used in Activity 2.2 could also be used to work out the distance to any object if you knew its size. However, under no circumstances should you try Activity 2.2 on the Sun because it would seriously damage your eyes.

Errors and uncertainties

The answer you obtained for the distance of the Moon in Activity 2.2 may have been rather different from the true value. This does not in any way reflect on your ability to perform the measurements but on unavoidable uncertainties resulting from the design of the experiment, such as the difficulty in judging when the coin exactly covered the disc of the Moon and the location of your eye in relation to the end of the rod. If your eye is not perfectly aligned with the end of the rod, or the coin is not placed in exactly the correct position, then you will introduce an **error** into your answer. There will also be an error introduced in your measurement of the diameter of a coin; for example if you used a ruler with millimetre divisions then the precision of the measurement is likely to be around half a mm. All these measurement errors are likely to combine to produce quite a large error in your final determination of the distance of the Moon. The error is the difference between your answer and the precisely correct answer. In some cases an error in a measurement may have important consequences (an example is the targeting of a spacecraft). In general, it is not possible to determine the error in a measurement. If you could then you would know the correct answer before you started and the measurement would not need to be made! However, it is important to have a good idea of the likely accuracy of any quantity or measurement. In the case of a spacecraft targeted to land on a small asteroid or comet, a knowledge of the positions of the spacecraft and target asteroid or comet to an accuracy within 100 metres would be fine, but if they were known only to the nearest 100 kilometres, the spacecraft may miss the target completely.

The **uncertainty** in a measurement or quantity is an estimate of the possible error that may be present. In Activity 2.2 you may have measured the position of the coin at 180 cm from your eye but may have thought any value between 175 and 185 cm was possible. You therefore specify the uncertainty as plus or minus five centimetres and express your measurement as 180 ± 5 cm.

Astronomy often requires observations and measurements that are at the limits of detection of the largest telescopes. A realistic estimate of the uncertainty in any measured quantity is vital if we wish to interpret the result. For example, the distances of the two galaxies that appear close to each other in the sky in Figure 2.3 may be measured as 6.9 and 8.0 megaparsecs (Mpc) respectively. We might conclude that they are separated by more than 1 Mpc (a vast distance compared with their dimensions) and are therefore just aligned along our line of sight by chance and cannot interact with each other. However, if the uncertainties are included for these measurements (i.e. 6.9 ± 1.0 Mpc and 8.0 ± 1.0 Mpc) then it is possible that they could both be at the same distance and we cannot rule out the possibility that they are interacting.

This module won't be concerned with the details of these uncertainties but it's important to recognise that many of the numbers that are quoted result from observations, so they are subject to uncertainties.

2.4.3 Angular sizes of astronomical objects

The distances to stars are so vast that almost all of them appear as points of light even when using the largest telescopes.

- If the Sun, with a diameter of 1.4 million km, were placed at the distance of the next nearest star, (1.3 pc) what would be its angular size?

  The Sun's angular size would be
  $57° \times (1.4 \times 10^6 \times 1000\ \text{m})/(1.3 \times 3.08 \times 10^{16}\ \text{m}) = 2.0 \times 10^{-6}$ degrees,
  which is $2.0 \times 10^{-6} \times 3600$ arcsec $= 7.2 \times 10^{-3}$ arcsec.

Such a small angular size cannot be measured with an ordinary telescope for two reasons. Firstly, the non-uniformity of the Earth's atmosphere causes stellar images to be smeared out to around an arcsecond in diameter. Secondly, even if this effect can be removed (for example by using a telescope in space), there is a theoretical limit to the angular size that can be seen. Larger telescopes can distinguish smaller angular sizes but it would require a telescope with a diameter of over 20 metres to just detect the Sun's disc at a distance of 1.3 pc. Using techniques of combining the light from telescopes to make in effect a much larger telescope, astronomers have been able to measure the angular diameters of some large and or nearby stars. The radii of most stars cannot be determined directly and are derived from other information (which we do not have space to describe here).

The angular *separations* of stars are much easier to measure and are particularly important for determining the orbits of binary stars (two stars in close orbit around each other) from which it is possible to derive the masses and other properties of the individual stars (see Chapter 5).

The distribution of stars in the sky is not uniform. Within our galaxy we can see **star clusters**, groups of hundreds or thousands of stars the closest of which have angular sizes of several degrees and are visible to the unaided eye (Figure 2.12 overleaf). They contain stars that formed at the same time and provide vital clues to our understanding of stellar evolution (see Chapter 5) The angular sizes of external galaxies range from several degrees (for the Magellanic Clouds, two companion galaxies to our own galaxy observable from the Southern Hemisphere as cloud-like patches separated from the Milky Way) to barely detected smudges in long exposure images taken with large telescopes. In principle the angular sizes of galaxies could be used to determine their distances, but the results are not sufficiently precise to be useful because of their very wide range of sizes.

Distances to astronomical objects are found using a variety of techniques, some of which are described in the next section.

2.5 Measuring distances

2.5.1 The distance ladder

There are many possible techniques for determination of distances to astronomical objects. Many of them use the concept of comparing an observed property of an object with the intrinsic value of that property, the difference being directly related to the distance of the object. You have already been introduced to the use of objects with known brightness (Section 1.5) or known size (Section 2.1). However, you have also learnt that there is no ideal object or property for determining distances throughout the Universe. An individual method may work well for a certain range of distances but be inaccurate or impossible to use outside this range. For example, a tape measure is suitable for measuring sizes of objects from centimetres to metres, but would be impractical for measuring the distance between towns.

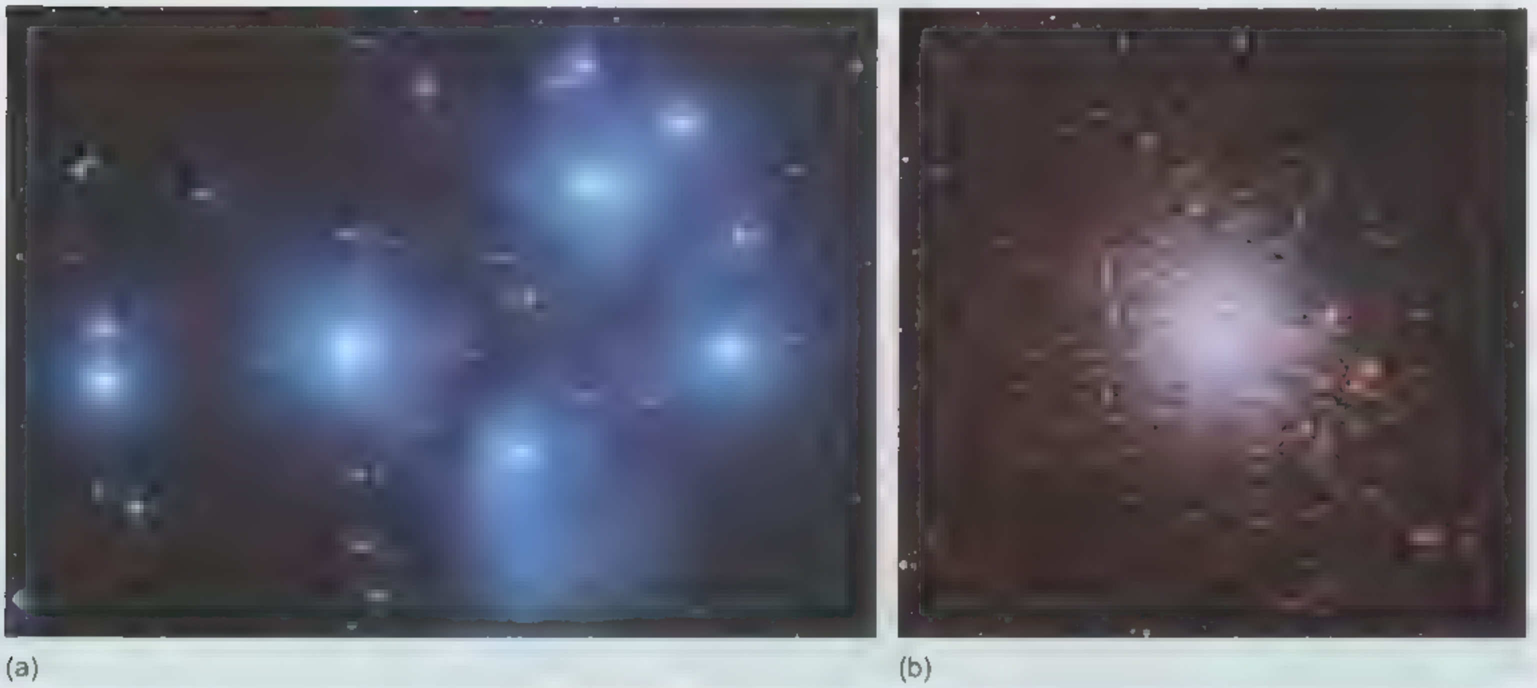


Figure 2.12 Star clusters in the Milky Way galaxy. (a) The Pleiades, a star cluster visible to the unaided eye in the Northern Hemisphere, has an angular size of almost 2°. It contains over 1000 stars and lies at a distance of about 130 pc. (b) The star cluster Omega Centauri has an angular size of around 0.5° and is visible to the unaided eye from the Southern Hemisphere. It lies at a distance of around 5000 pc and contains several million stars.

Astronomers use many different methods to determine distances, from the scale of the Solar System to the most distant galaxies. Most of these methods involve the use of another method to support them, leading to the concept of the **distance ladder**. It is so called because the top steps (the methods to determine the largest distances) can only be reached by having the lower steps already in place. Some methods (or steps) only provide *relative* measurements of distance (i.e. that object B is twice the distance of object A) and require another method to derive absolute distance calibration. In this example, the step below on the ladder can give the distance of object A but not B. The combination of the two methods gives the distances of both A and B.

The first steps on the distance ladder give the size of the Earth, the distance from the Earth to the Sun (the astronomical unit, AU) and then to the nearest stars.

2.5.2 Parallax: distances to nearby stars

Humans, and many animals, use parallax to provide depth perception and estimate distances to nearby objects. This is because each eye provides a slightly different view of the position of a nearby object compared with the background view (see Figure 2.13). Astronomers use the same technique to determine the distances to nearby stars (see Figure 2.14). The angle p, called the parallax is greatly exaggerated in these views and is less than 1 arcsecond for even the closest star. The unit of distance, the parsec, is defined as the distance of a star which has a *par*allax of 1 arc*sec*ond.

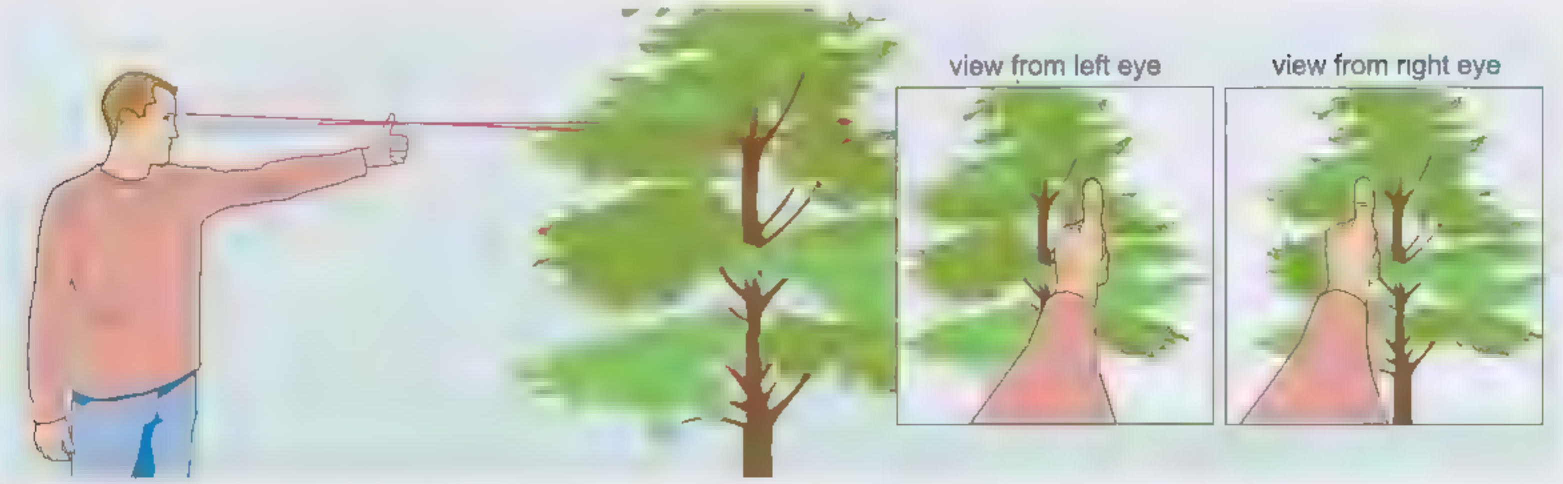


Figure 2.13 The apparent change in position of a nearby object seen with left and right eyes.

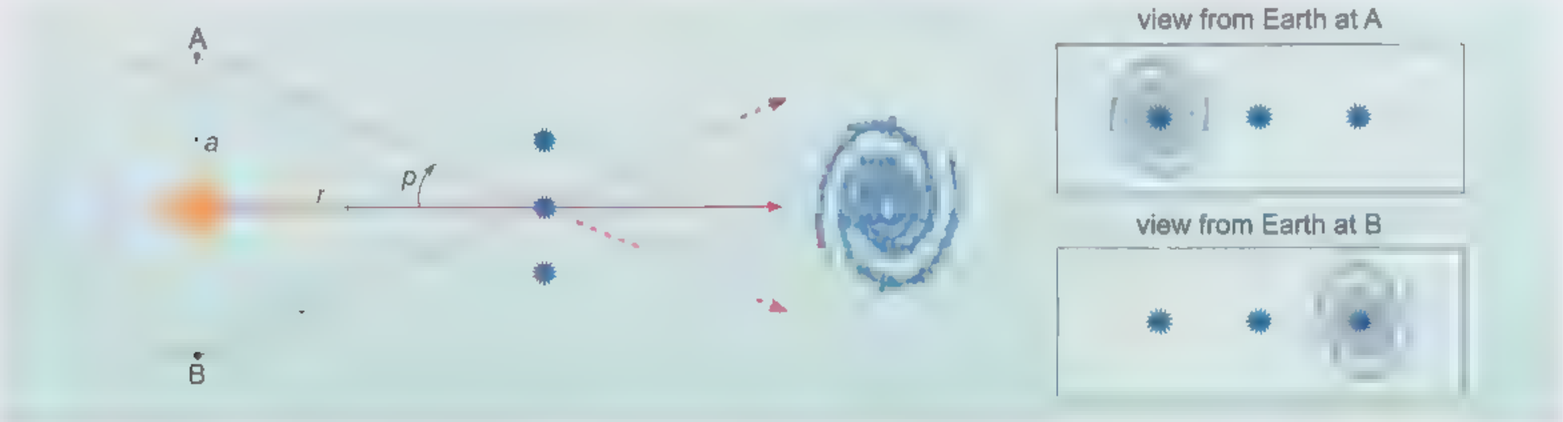


Figure 2.14 The apparent change in position of nearby stars observed six months apart from the Earth on opposite sides of its orbit around the Sun. The angle p is the parallax. The distance r can be calculated if the angle p is measured and the distance a is known (1 AU).

Clearly, very precise measurements are required to determine the parallax, and until the advent of space telescopes, it was only possible to determine the distances of very nearby stars (to distances of around 100 pc). Despite this, the technique provided the bridge between distance measurements in the Solar System and those of the stars.

Significant figures

If you are told that the mass of the Earth is 6×10^{24} kg, what does that mean? It does not necessarily mean that the mass is exactly 6×10^{24} kg, but that it lies closer to 6×10^{24} kg than it does to either 5×10^{24} kg or 7×10^{24} kg. If it had been written as 6.0×10^{24} kg, you would be justified in assuming the mass is nearer to 6.0×10^{24} kg than 5.9×10^{24} kg or 6.1×10^{24} kg. A value such as 6 (or six times any power of ten) is said to be 'quoted to one significant figure', whereas 6.0 is 'quoted to two significant figures' and implies greater precision. In fact, the mass of the Earth is known to four significant figures (5.974×10^{24} kg). Expressing the mass of the Earth as any of 6×10^{24} kg, 6.0×10^{24} kg or 5.97×10^{24} kg is correct to the number of significant figures quoted.

In the discussion of errors and uncertainties in Section 2.4.2, the distance to one of the two galaxies in Figure 2.3 was quoted as 6.9 Mpc with an uncertainty of 1.0 Mpc. From this, the true distance is probably somewhere in the range of 5.9 to 7.9 Mpc (although it may be a little outside this range since the uncertainty is an estimated quantity).

■ Is it reasonable to quote the distance to the galaxy as 6.921 ± 1.0 Mpc or 6.9 ± 1.023 Mpc?

□ Neither way of stating the distance is correct. The first quotes the distance to 4 significant figures and implies that there is a significant difference between a derived distance of 6.920 and 6.921 Mpc. The quoted uncertainty is inconsistent with this. In the second case, the uncertainty is quoted to too many figures. Since it is an estimate, it should be quoted to only one, or at most two, significant figures.

A common error made by students using electronic calculators is to quote an answer to a calculation to the number of figures given on the display. If you were asked to calculate how much further away the second galaxy is than the first, your calculator will tell you $8.0 \div 6.9 = 1.159\,420\,29$. However, if the quoted uncertainties are a good guide, it could lie anywhere between $9.0 \div 5.9 = 1.5$ and $7.0 \div 7.9 = 0.89$. The best way to state the ratio is therefore as 'the ratio of distances is about 1' or 'they are about the same distance away': more than two significant figures implies a false level of accuracy.

2.5.3 Stellar motions

The patterns of stars on the sky appear to be fixed. The most prominent constellations (patterns of stars) that are seen today were mapped over three thousand years ago. However, the stars do move relative to each other at speeds of tens to hundreds of kilometres per second. It is only because they are so far away that the apparent change in position is so small; the apparent changes in position of stars are measured in units of milli-arcseconds (an angular change in position of a thousandth of an arcsecond) per year! This is the motion of the star projected onto the plane of the sky and is given the rather confusing name *proper motion*.

It may surprise you to learn that it is much easier for astronomers to measure the movement along the line of sight to the star. This is because they can take advantage of the wave properties of light and a phenomenon known as the **Doppler effect**. Figure 2.15 illustrates the Doppler effect occurring with sound waves in an everyday situation.

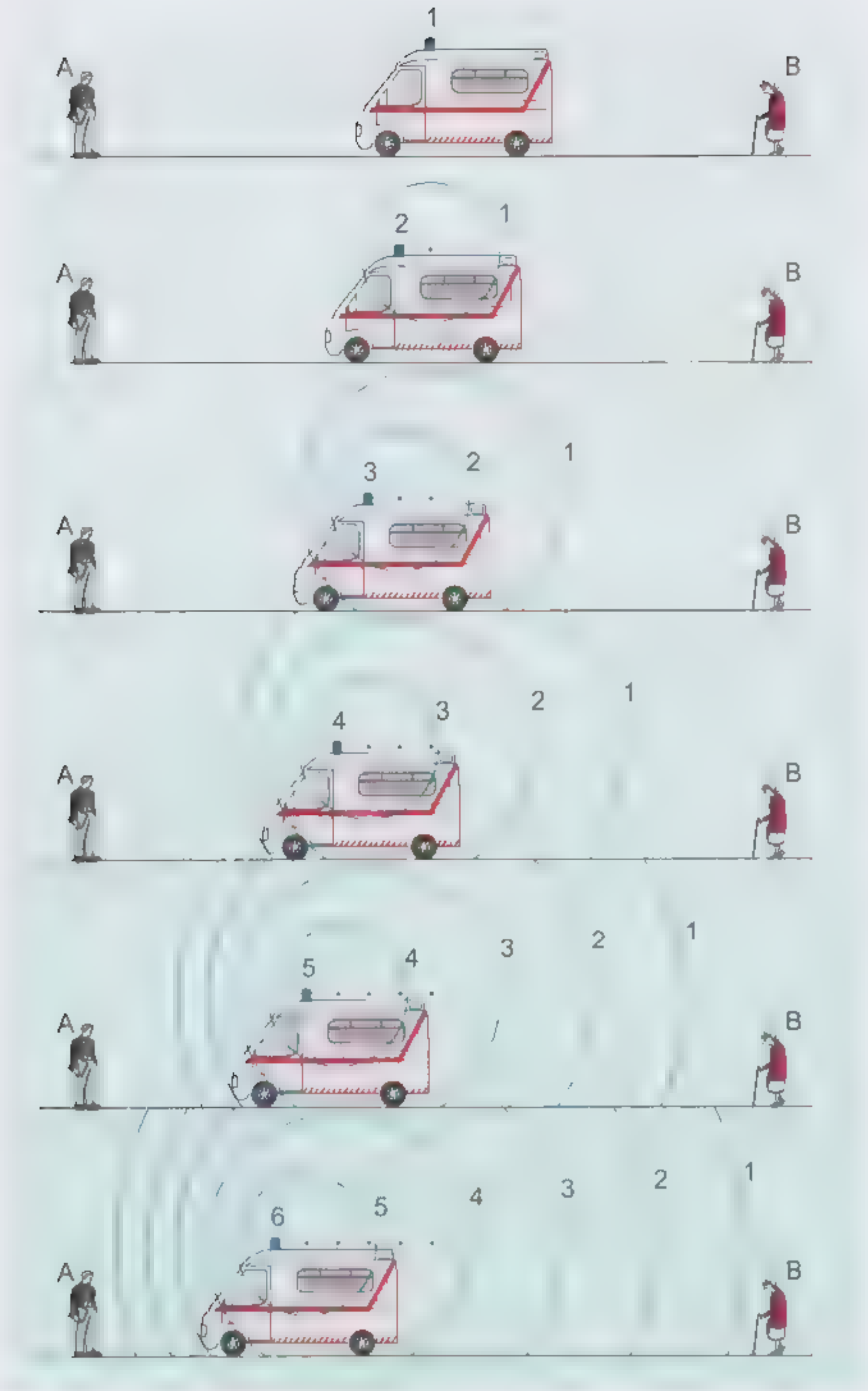


Figure 2.15 A demonstration of the Doppler effect w.th sound (not to scale).

The ambulance sounds its siren as it moves towards observer A. Six successive time intervals are shown in the six sketches, with the curved lines representing successive crests of the sound wave emitted by the siren. You may like to think of the curved lines as being similar to the ripples produced when a stone is dropped into a pond – the wave crests spread out from the centre just as shown here. The wavelength is then just the distance between any two successive crests at any point. By the time the second wave crest is emitted, the ambulance has caught up slightly with the first wave crest. By the time the third wave crest is emitted, the ambulance has caught up with the second wave crest, and so on. The consequence is that a person at A will perceive a sound wave with a shorter wavelength than that emitted by the siren when at rest, while a person at B will perceive a longer wavelength.

The same process occurs with light. If a star is moving towards the observer the wavelengths of absorption lines in the spectrum of the star are shorter than expected (they are said to be blueshifted) and if the star moves away from the observer the wavelengths are longer than expected (redshifted). The wavelength shift can be used to calculate the speed of the star along the line of sight, called the **radial velocity**. A wavelength shift of 1 Å at optical wavelengths corresponds to a speed of around 50 km s^{-1}.

Radial speeds determined using the Doppler effect are important for determination of masses of stars (in binary systems – see Chapter 5) and planets in extrasolar planetary systems (see Chapter 7).

2.5.4 Standard candles

The intrinsic brightness of an object is called its **luminosity**, defined as the rate at which energy is carried away by electromagnetic radiation. The luminosity of the Sun ($L_{\odot}$) is a staggering 3.8×10^{26} watts, but the Sun is only a very average star.

If the luminosity of a type of astronomical object is known, then its apparent brightness can be measured and its distance determined using the inverse square law (see Section 1.5). The luminosity of a star is a measure of its total power output. The apparent brightness of the star, as seen by an observer on the Earth, depends on both its luminosity and its distance. If the luminosity is known independently, then the distance can be calculated.

Objects that can be used for such distance determinations are called **standard candles**.

■ What properties are required for an object to be a standard candle?

The object must have a known luminosity, be intrinsically bright so that it can be seen over large distances, it must have properties that allow it to be recognised and it must be commonly found in locations whose distance we wish to measure.

The luminosity of a standard candle is usually derived for a nearby example using an independent method, and the assumption is made that all other objects with the same properties have the same luminosity. If a star like the Sun was used as a standard candle it could be detected up to distances approaching 1 megaparsec. Although this encompasses the whole of the Milky Way and some nearby galaxies, solar-type stars are not very useful as standard candles because they are not easily distinguished from other types of stars with the same temperature but different luminosities.

An example of a standard candle that is widely used is a common type of variable star called a **Cepheid variable**. These stars are sufficiently bright (from 100 to 10000$L_{\odot}$) that they can be identified in all parts of our galaxy and in other nearby galaxies and galaxy clusters They are easily identified by the characteristic way that they change in brightness and their luminosity is related to the period of that variation. These characteristics were recognised by a study of Cepheid variable stars in the Small Magellanic Cloud over a century ago.

Another example, that has even higher luminosity, and can therefore be detected at much greater distances, is a supernova. Although very rare, these catastrophic stellar explosions (see Chapter 6) can often outshine their host galaxy for a short time Certain types of supernovae can be used as standard candles beyond the range of Cepheid variables.

The combined light from an entire galaxy can be used as a standard candle to determine the distance of a cluster of galaxies. The brightness of galaxies varies enormously, so the success of this method requires careful choice of the type of galaxy. One approach is to make a statistical argument, assuming the distribution of galaxies in similar size clusters is the same so that, for example, the third brightest galaxy in a cluster has the same brightness as the third brightest in another similar cluster.

Standard candles allow us to determine distances exceeding 1000 Mpc, approaching the limits of the observable Universe. In the next two chapters we will investigate the limits of the Universe and our current understanding of its origin and evolution.

End-of-chapter questions

Question 2.1 Jupiter has a diameter of 140 thousand kilometres and orbits at 778 million kilometres from the Sun. Write the diameter and distance of Jupiter in metres using scientific notation.

Question 2.2 How many times more energy does an X-ray photon with a wavelength of 1 ångström have compared with a radio photon with a wavelength of 1 m?

Question 2.3 Use Figure 2.7 and Table 2.4 to explain how the colour of a B-class star is likely to compare with that of an M-class star. How would it be possible for the M-class star to appear brighter than the B-class star if they were at the same distance from the observer?

Question 2.4 A large sunspot can be several times the size of the Earth. What is the angular size of a sunspot that has the same diameter as the Earth? Express your answer in arcseconds. What fraction of the Sun's diameter is the sunspot?

Question 2.5 What is wrong with the following representation of distances?

(i) 34.135 ± 3 ly; (ii) 29.35 ± 6.14 pc; (iii) 6 ± 0.01 AU.

Now go to the module website and do the remaining activities associated with Chapter 2.

Activity 2.3 - 15 mins
Activity 2.4 - 20 mins
Activity 2.5 - 10 mins

Chapter 3 The Universe at large

3.1 Introduction

Perhaps one of the most astonishing things about the Universe is that it's possible for us to build telescopes that can see nearly all the way back to the Big Bang. Light doesn't travel infinitely quickly, so the further away a thing is, the longer it takes the light to get to us. The light from the most distant things seen by telescopes took most of the lifetime of the Universe to get to us.

By comparing the distant, early Universe to the nearby, present-day Universe, astronomers can see how things in the Universe gradually take shape. This chapter takes you on a brief tour back in time from today almost to the start of the Universe.

3.2 The Milky Way and its neighbours

If you have a dark enough night sky, you'll be able to see the Milky Way. It's a hazy band of light right across the sky, about the width of your hand at arm's length. It's easiest to see away from city lights and during a new Moon. Ancient astronomers speculated over what the Milky Way is made of, but Galileo made a huge breakthrough, which he published in 1610. He'd turned his telescope to the Milky Way and found it was made of many faint stars, too faint and too crowded to be distinguished with the naked eye.

But why are the stars seen as a band across the sky, and not some other pattern? Immanuel Kant speculated in 1755 that it was because the Earth is inside a huge rotating disc of stars held together by gravity. This is broadly the picture today of what a galaxy is. Even in 1755 astronomers were aware of 'spiral nebulae', which Kant speculated were other galaxies that are now called spiral galaxies. You can even see one nearly edge-on spiral galaxy with the naked eye as a faint fuzz: the Andromeda galaxy (Figures 3.1 and 3.2), also known as M31. The M refers to the Messier catalogue of nebulae, published in 1774. (A nebula is a fuzzy luminous patch in the sky, and the plural is *nebulae*.) The Andromeda galaxy is 2.5 million light-years away, i.e. the light has taken 2.5 million years to reach us, yet you can see it with the naked eye!

How far away is the Andromeda galaxy?

One light-year is about 0.3 parsecs, so 2.5 million light-years is about (2.5 × 0.3) million parsecs, or about 0.8 million parsecs. This is 800 000 parsecs, or 800 kiloparsecs (800 kpc for short). One light-year is also about 10^{16} metres, so the Andromeda galaxy is about 2.5 million × 10^{16} metres away, or $2.5 \times 10^{6} \times 10^{16}$ metres or 2.5×10^{22} metres away.

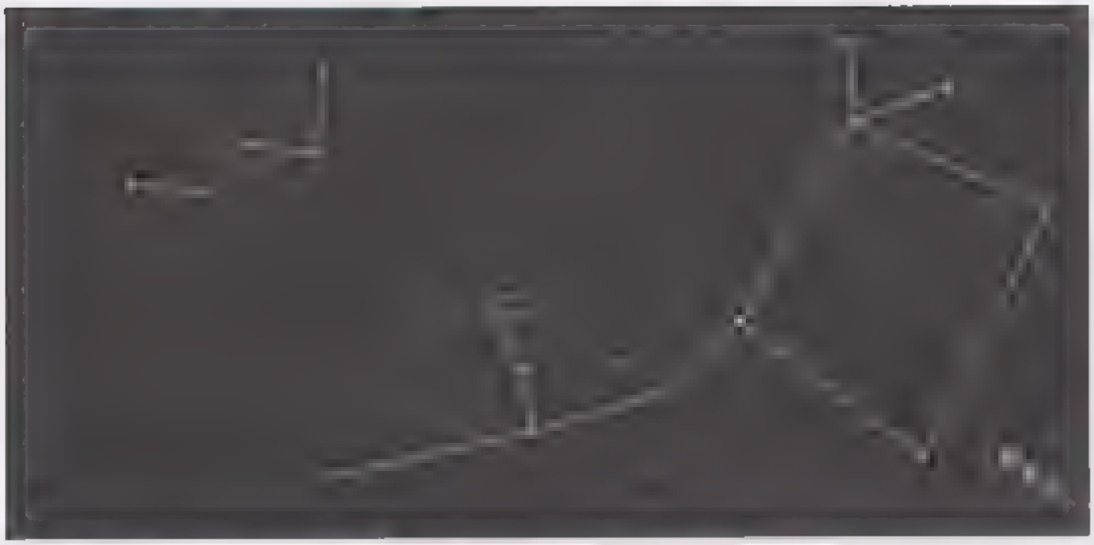


Figure 3.1 How to find the Andromeda galaxy. This picture is about one-third of the horizon wide (i.e. about 120°), and the orientation will depend on when and from where you're looking. The picture is slightly distorted to fit on this flat page. The W of Cassiopeia is at the top left, and the square of Pegasus is to the right. These constellations aren't visible at night from everywhere and at any time. Andromeda is most easily seen from the Northern Hemisphere. The Andromeda galaxy is the ellipse, and is often just visible to the naked eye as a faint smudge. A good trick to see a faint object is to look slightly to one side of it, because your eye has more brightness sensitivity (and less colour sensitivity) away from the centre of your vision.

Kant called these galaxies 'island universes', and that terminology persisted right into the early twentieth century. However, the meaning of the word 'universe' has since drifted and it's now taken to mean everything that exists, galaxies included. Unfortunately the meaning of 'universe' is still drifting, with some theoretical physicists speaking of 'other universes' in some speculative contexts, but this book will stick to using 'the Universe' to mean everything that exists.

Figure 3.2 The Andromeda galaxy, seen through a telescope.

Kant's suggestion that 'spiral nebulae' were external galaxies was intensively debated among astronomers, but was only proved as recently as the 1920s. In 1922 the astronomer Ernst Opik estimated the distance to the Andromeda galaxy, finding it to be clearly far more distant than the stars in our galaxy. Then in 1929, Edwin Hubble detected individual stars in Andromeda. By identifying particular types of stars with a characteristic luminosity, then measuring how bright they appear, he confirmed the Andromeda 'nebula' is a distant object outside our galaxy. Such objects are called **extragalactic**.

Astronomers often refer to our galaxy (the Milky Way) as simply the Galaxy, with a capital G. The Galaxy is about 10^5 light-years in diameter. It is a spiral galaxy and most of its stars are in a **disc** that's about a thousand light-years in thickness. There are about a hundred billion stars in the Galaxy (i.e. 10^{11} stars). There's a **bulge** in the centre of the Galaxy, and the stars in that region tend to be older and redder than the rest. There's also a much fainter, roughly spherical **stellar halo** encompassing the disc. Figure 3.3 shows the structure of the Galaxy, and what it might look like face-on.

But where have these halo stars come from? Some of the answers to this will be discussed shortly, but first there's a giant component of the Galaxy that has been missed out – in fact, it has more mass than all the stars and all the gas put together.

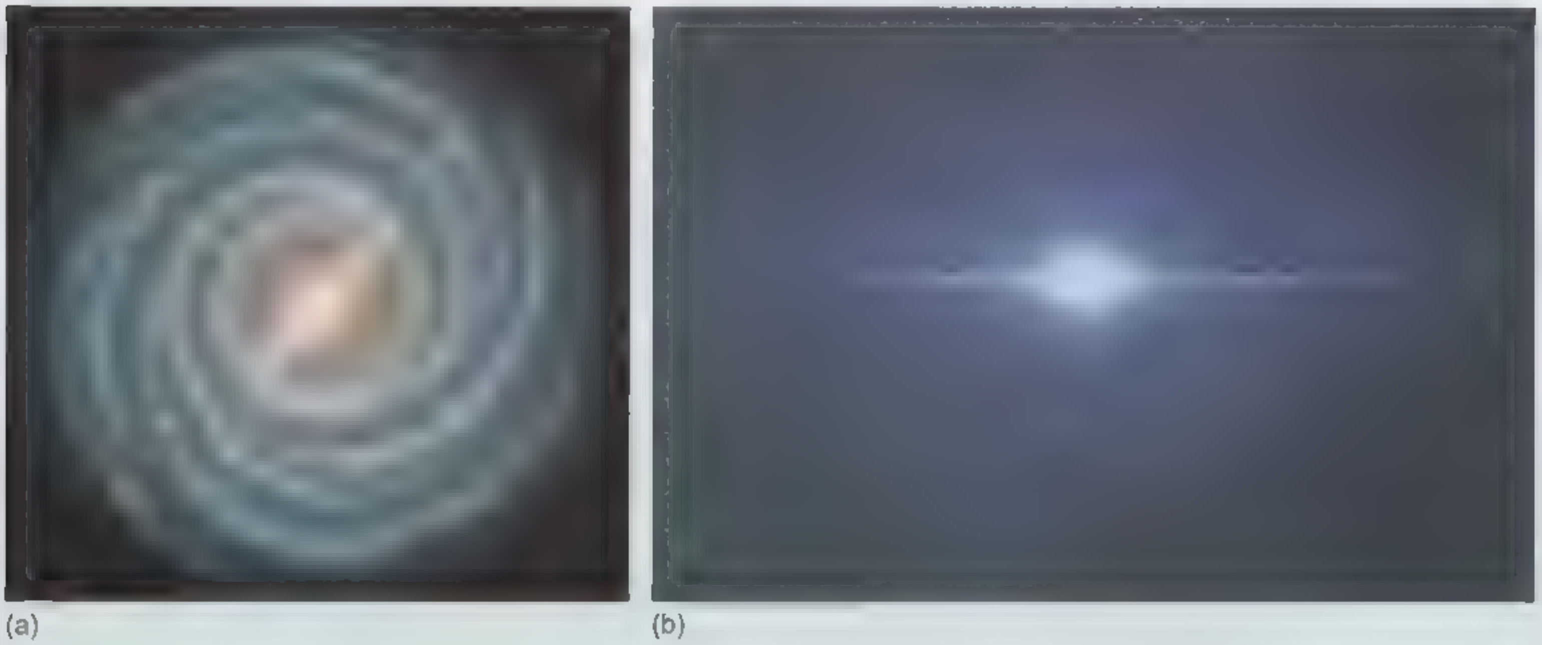


Figure 3.3 An artist's impression of our galaxy, the Galaxy, also known as the Milky Way, seen (a) face-on and (b) edge-on.

- The more mass that a galaxy contains, the faster its stars go in their orbits. Why?

  The more powerful the tug of gravity, the faster the stars have to move to avoid falling towards the centre.

Many spiral galaxies have had their masses estimated from how fast their stars are orbiting. The trouble is, if you add up all the masses of the stars themselves and throw in the mass of the galaxy's gas too (it turns out that the gas can be detected with radio telescopes), it doesn't come close to those estimates from stars' orbits. If the stars and gas were all there is, the stars would be flung out of the galaxy, because they're moving so quickly. So what's keeping galaxies together? It's something non-luminous because it cannot be seen. Astronomers have hypothesised **dark matter** to explain this.

Dark matter is also needed to explain the motion of galaxies in galaxy clusters (which you will meet later in this chapter), the motion of galaxies in general, and other data. The astronomical evidence for some extra component or effect is now very strong. Most astronomers accept this as evidence for dark matter and use it as the most plausible working hypothesis (no scientist has *beliefs* about a scientific theory, at least in the ideal world), though a few have preferred to suppose that Einstein's theory of gravity is wrong. Many theories of sub-atomic particles (protons, neutrons and a whole panoply of more exotic things) predict the existence of various sorts of dark matter (it may be detectable at the Large Hadron Collider and in other experiments), and it's proved quite difficult to make a new gravity theory that explains the multitude of observations without dark matter. This book works on the assumption that dark matter probably exists.

Dark matter is perhaps badly named. If you imagine a bucket of dark matter, you would probably think of a container of something like black sludge. In fact dark matter must be transparent, or it wouldn't be possible to see galaxies at all. Dark matter makes up over 80% of all the matter in the Universe, yet at the time of writing (2012) scientists have no idea what dark matter is. In terms of the fundamental particles (those that do not have a substructure), physicists do know that it's *not* mainly baryons (baryons include the protons and neutrons in visible matter), because if it were, there would have been many more nuclear reactions in the very early Universe. These would have left a different pattern of element abundances throughout the Universe.

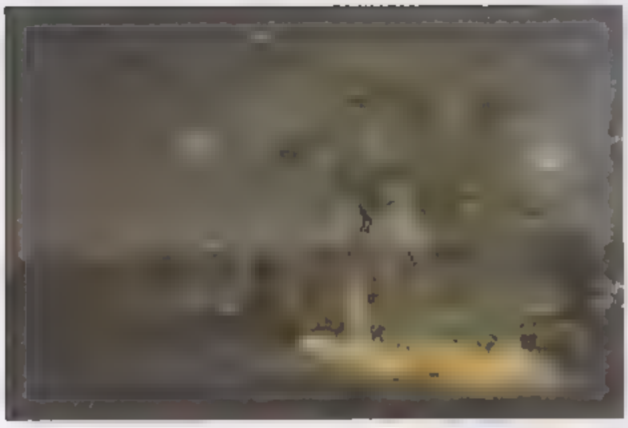


Figure 3.4 The Large and Small Magellanic Clouds, as seen from New South Wales, Australia.

Our galaxy has neighbouring galaxies that it's slowly eating. The Large Magellanic Cloud and the Small Magellanic Cloud (Figure 3.4), which can often be seen with the naked eye from the Southern Hemisphere, are dwarf galaxies in the process of being accreted by the Galaxy. This is not at all unusual. Galaxy accretion and mergers turn out to have been very common in the past histories of present-day galaxies. These accretion events are the sources of some of the stars in the stellar halo of the Galaxy. The stellar halo also contains stars from early on in the formation of the Galaxy. Our galaxy is itself falling towards the Andromeda galaxy, and in about five billion years time, these galaxies will merge. What will this be like? For one thing, the stars themselves won't collide – they are too sparsely distributed. The galaxies will splash together, flinging some stars out in the process. Figure 3.5 shows an example of two spiral galaxies caught in the act of merging. Eventually the combined system will settle down, perhaps resembling an elliptical galaxy. Scientists have calculated that there's a 12% chance that our own Solar System will be ejected during this great merger. This wouldn't affect life here as we know it because the Earth will have long since become uninhabitable. About a billion years from now, the Sun's luminosity will be too high for liquid water to exist on Earth. At that time, the Andromeda galaxy will loom as large in the sky as either Magellanic Cloud does today.

3.3 The Local Group and galaxy clusters

The impression you may have had from the last section is that our near neighbourhood is quite dynamic. This dynamic environment isn't at all unusual. The rest of this chapter will take you on a tour of the wider Universe. At each stage, as you're introduced to larger and larger structures, keep in mind that the Universe still appears dynamic. Galaxies are flowing towards regions of higher density, because of gravitational attraction. This also means galaxies flow out of regions of lower density. As time goes on, the higher density regions attract more and more galaxies, and the under-dense regions become less and less populated. The bigger the scale, the longer it takes galaxies to fall in, but it's still always a dynamic picture.

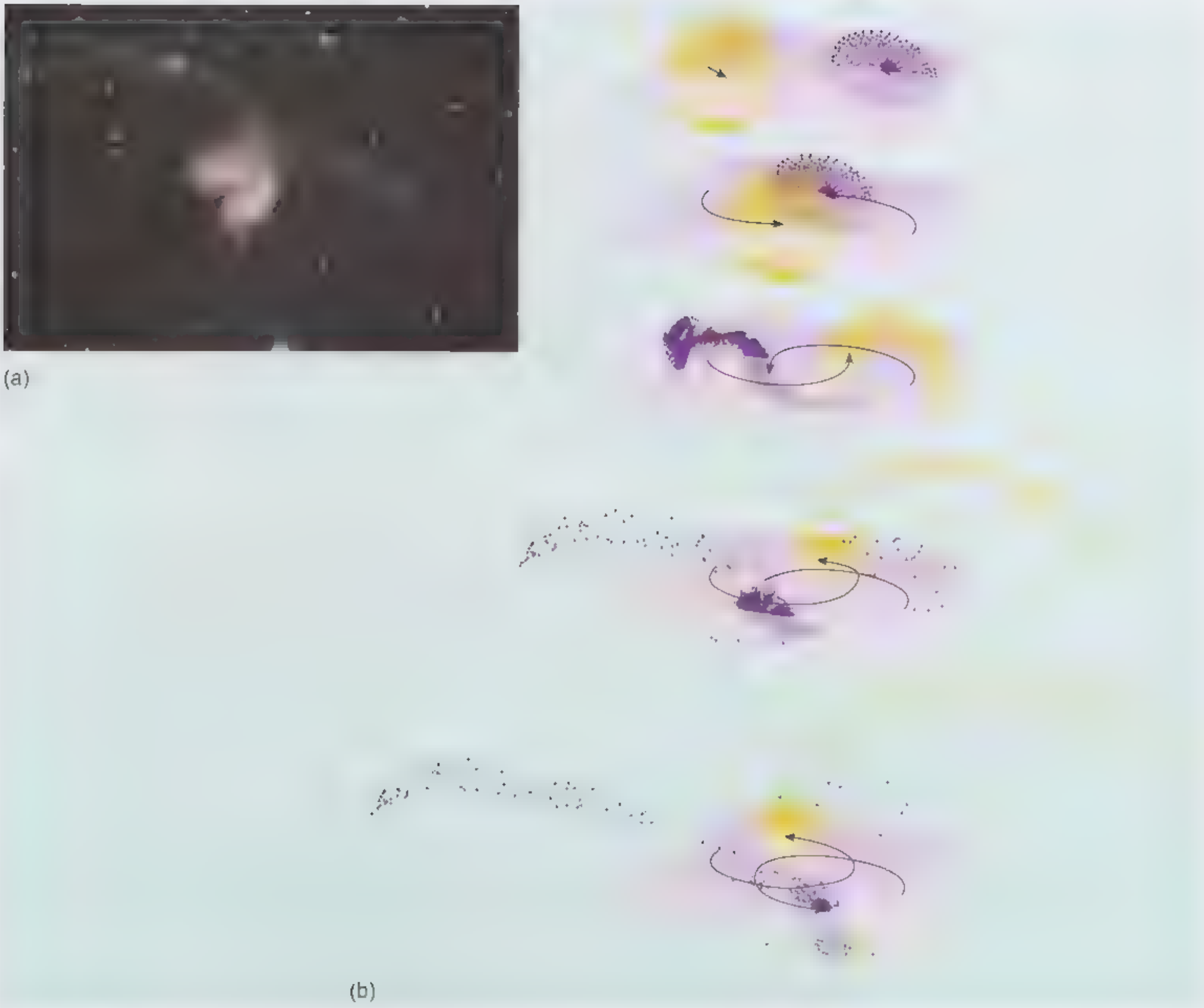


Figure 3.5 (a) The Antennae galaxies, which are in the process of merging. Note the streams of stars flung out in the process of the merger. (b) A computer simulation of the Antennae galaxies as a merger of two spirals. Time is running from top to bottom in this sequence. Chapter 6 will show you this galaxy collision in more detail.

The Magellanic Clouds are not the only galaxies near our Milky Way galaxy, though they're the best known. An even closer neighbour, the Sagittarius dwarf elliptical galaxy, was only discovered in 1994. The reason it hadn't been spotted before is that from our perspective, it's behind the bulge of our own galaxy. The closest galaxy to our own galaxy discovered so far (early 2012) is the Canis Major dwarf galaxy; at a distance of 7.6 kpc it's closer to Earth than the bulge of our own galaxy. From our point of view it's behind the Milky Way, making it very hard to detect – it was only discovered in 2003. There are several other dwarf galaxies in the neighbourhood of the Galaxy.

Moving slightly outwards, you meet the Andromeda galaxy. As you've seen, our Milky Way system is falling towards it. Moving out further, our galaxy is in an over-dense region of galaxies compared to the cosmic average. This is known as the **Local Group** of galaxies. There are over 40 galaxies in the Local Group, including the Milky Way and Andromeda. Moving further out again, there are more groups of galaxies. The galaxies in galaxy groups are gravitationally bound together.

Stepping out a little further still, there is a much larger gravitationally-bound structure: the Virgo cluster of galaxies, 16.5 million parsecs away (16.5 megaparsecs, or 16.5 Mpc for short). It contains over 2000 galaxies and even more visible matter in the form of hot gas. It's too faint to be seen with the naked eye, but its apparent diameter on the sky is ten times bigger than the full Moon! Its mass is enormous, at over 10^{15} times the mass of our Sun. Galaxies and groups of galaxies on this side of the Virgo cluster are falling towards it through its gravitational attraction, and will eventually join the cluster.

Galaxy clusters are the largest gravitationally-bound structures in the Universe. Just like the stars in a spiral galaxy, the orbits of galaxies in galaxy clusters give strong evidence for dark matter. In fact, this was the very first line of evidence made for dark matter. In galaxy clusters **elliptical galaxies** are much more common than spiral galaxies. The stars in elliptical galaxies are not confined to a disc but rather orbit at all inclinations. The stars tend to be older than in spiral discs (much like the bulges of spiral galaxies), and they tend to have little ongoing star formation (at least today). It seems that clusters tend to have many more elliptical galaxies compared to regions outside galaxy clusters. Perhaps if a spiral galaxy passes through the hot gas inside a cluster, this gas helps strip out the gas from spiral discs. Also, the many close encounters from the (relatively) crowded environment of a galaxy cluster can disrupt the motions of the gas and stars in the spiral discs.

Now that you have met spiral and elliptical galaxies, you need a more complete list of the types of galaxies that have been found in the Universe. Elliptical galaxies can be classified in terms of how long and thin they appear, and spiral galaxies by how tightly wound the spiral arms appear. Also, spiral galaxies can have bars in the centre – in fact, it appears likely that our own galaxy is a barred spiral. Finally, there are the irregular galaxies that have no clear common identifying features. Edwin Hubble, who resolved the individual stars in the Andromeda galaxy and confirmed its extreme distance, went on to study other galaxies. He proposed a 'tuning fork' classification that is still in use today. This is shown in Figure 3.6.

Now, there's a terribly misleading terminology in use by astronomers that you need to be aware of. Elliptical galaxies are sometimes called 'early type', and spirals are sometimes called 'late type'. These 'early' and 'late' names go back to Hubble's original classification, at a time when galaxy evolution was not at all well-understood. Astronomers now know that there is *no* such evolutionary sequence from what are called 'early-type galaxies' to what are called 'late-type galaxies', but astronomers have stuck with the names nonetheless. Do not let these misleading names confuse you.

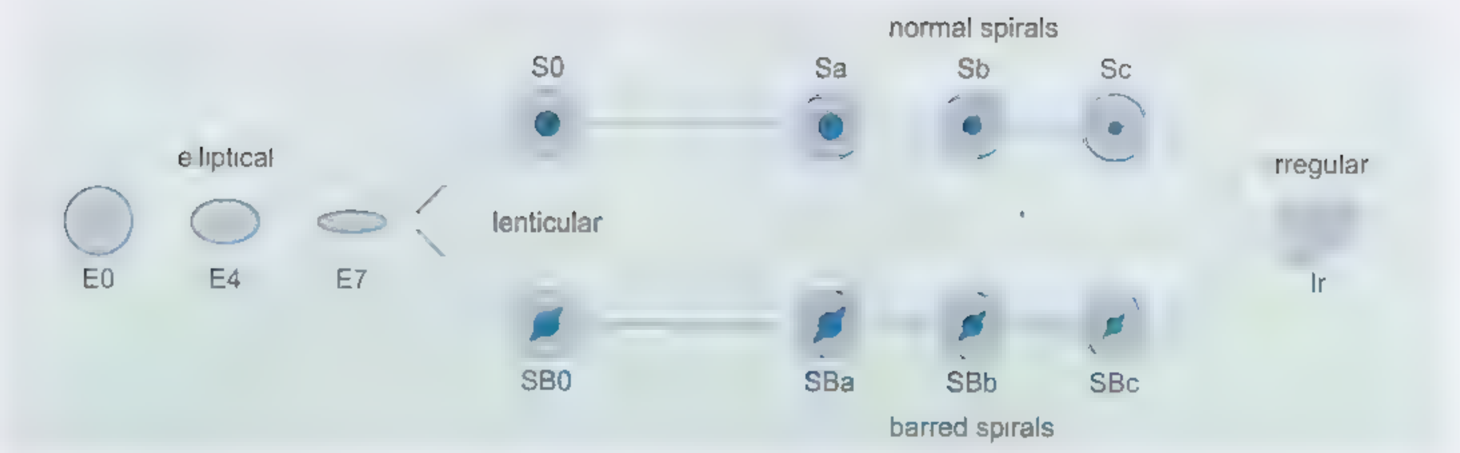


Figure 3.6 The Hubble tuning fork classification scheme for galaxies. The elliptical galaxy types E0, E4 and E7 are shown, but there are also other intermediate shapes such as E2 or E6. The Milky Way is somewhere on the barred spirals fork, but researchers are not completely sure where!

Interpreting pictures of three-dimensional objects

We are all accustomed to looking at pictures of three-dimensional objects, such as people's heads, and forming a clear idea about what the whole object looks like even though the evidence we have only relates to a single viewpoint. This is a useful skill that is partly built on our past experience of looking at similar objects. However, when dealing with scientific images, it is important to extract as much information as possible from an image, but not to read into it more than is there. Speculation about what might be in an image is not necessarily bad, but you must be clear about the point at which observation stops and speculation begins.

Examining pictures of galaxies is fertile ground for developing several interpretive skills. For example, what shape are elliptical galaxies? A particular image may have an elliptical outline, and may be classified as being somewhere in the range E0 to E7, according to the relative lengths of the long and the short axes of the observed ellipse (Figure 3.6), but what would such a galaxy look like if it was rotated through 90°? What are the true three-dimensional shapes of elliptical galaxies and to what extent can they be discerned from two-dimensional images?

Try a little experiment. Pick up a pen and look at it end-on. What shape do you see? If it's a ball point pen, probably a circle, rather like an E0 galaxy. Now look at the pen sideways on. What kind of elliptical galaxy does it most resemble? It may not look much like any elliptical galaxy, but it's probably closest to a cigar-shaped E7 galaxy. Hold the pen horizontally, in front of your eyes and rotate it from the sideways view to the end-on view. Can you see it pass through the stage where it looks like an E4? An E0 galaxy has a circular appearance. Does that mean it is really a spherical gathering of stars, or might it be an end-on view of a cigar-shaped distribution?

A pen is a solid object and light is reflected from its surface, which makes it easier to interpret the pen as a three-dimensional object rather than an outline. Elliptical galaxies are different. They emit light rather than reflect it. It is possible to work out the three-dimensional shape of some elliptical galaxies, but doing so generally requires detailed studies of the movements of stars within that galaxy, together with some assumptions about the way those motions influence the overall form of the galaxy. On the basis of detailed studies, astronomers have evidence that some elliptical galaxies are ellipsoidal. Such galaxies will have an elliptical appearance from every direction but the detailed form of the ellipse may change with the direction of view. A full appreciation of the shape of an elliptical galaxy is unlikely from any single image of that galaxy.

Similar comments apply to spiral and lenticular (i.e. lens-shaped) galaxies, which are generally seen as inclined discs. In some cases these may even be difficult to distinguish from elliptical galaxies!

Always keep in mind the distinction between the true shape of an object, and the shape it appears to have from a particular viewpoint, especially if it is illuminated in a way that is unfamiliar.

3.4 Dark matter

So far, the picture painted of the Universe is of many galaxies, some merging with each other in huge collisions that throw off stars and some being pulled towards each other, making bigger and bigger gravitationally-bound structures. What if you look at this dynamic, merging Universe from the point of view of the dark matter instead of the galaxies?

From the point of view of the dark matter, the Universe looks much simpler. Dark matter particles are usually treated by theoretical astrophysicists as 'collisionless' particles, meaning that dark matter particles don't bounce off each other like gas particles. Instead, they pass right through each other. Scientists don't yet know quite why, because it's not yet understood what the dark matter particles are, but it appears to fit the available data.

This makes a 'gas' of dark matter particles quite strange. It doesn't have any turbulence or viscosity or air resistance. In fact, right now as you are reading this, and our Solar System is orbiting the centre of the Galaxy and passing though the Galaxy's dark matter halo, dark matter particles are streaming through your body – and in fact the whole Earth – without you noticing it.

Dark matter particles also don't absorb or emit light. The only thing that appears to be happening with dark matter is its gravitational attraction to itself and to other matter. This means it can form transparent gravitationally-bound clumps, but they won't emit light or have nuclear reactions like stars, nor can dark matter form the discs of material that you will meet later in Chapter 6 when the formation of stars and of planetary systems is discussed. Dark matter clumps are therefore just left in a fairly spherical state, which is why the halo of the Galaxy is fairly spherical, as are the haloes of nearly all galaxies.

The dynamic history of the Universe also looks simpler from the point of view of the dark matter. Dark matter haloes merge with other dark matter haloes to form bigger dark matter haloes. The only relevant physics is the law of gravity. Haloes move out of under-dense regions towards bigger clumps by the force of gravity, and the Universe is a continual process of dark matter haloes merging and making bigger haloes. Large dark matter haloes look just like scaled-up smaller ones – there are no distinguishing features with fixed characteristic sizes. It's a bit like having the Universe scattered with billions of peas of different sizes – except that the haloes are a lot more alike than real peas, because real peas have skin and cells with very characteristic thicknesses and sizes, whereas dark matter haloes don't have features with characteristic sizes.

- Can you have dark matter haloes on their own, without any galaxies in them?

  Yes, but they're typically much harder to detect. Theoretical models predict that the haloes that lack galaxies today are almost exclusively the least massive haloes, with masses of about 10^{11} times the mass of our Sun or less.

So why does a galaxy cluster, which has one big dark matter halo, not have one great big galaxy in it, instead of thousands of smaller ones? To answer this, you need to switch back to looking at visible matter, where things are once again much more complicated. The visible matter can exist in many different states: hydrogen gas in various states (lone atoms, molecules, sometimes with electrons missing), other atoms such as helium, other molecules, dust, stars, planets, and so on. The processes that go on between all these different states are complicated, and (it turns out) they can have many characteristic size scales and timescales. For example, stars have characteristic lifetimes, sizes and life histories, which you will meet later in this book. These characteristics follow ultimately, and in a very non-trivial way, from considerations of atomic nuclei. Many properties of everyday objects, from salt crystals to nylon fibres to copper wires to snowflakes, follow ultimately from atomic or molecular microscopic properties (including the ability to make crystals). The tremendous complexity of the visible Universe is to a large degree due to the ability to make atoms and molecules. Collisionless dark matter particles can't form atoms or molecules, which is why dark matter haloes are so much simpler than visible matter. As a result, a large halo of collisionless dark matter must look just like a scaled-up smaller one, and this is why there are no distinguishing features that have characteristic sizes.

This is one way of explaining why an entire galaxy cluster doesn't look like one giant galaxy, but instead is comprised of many smaller galaxies and lots of hot gas (plus of course the roughly spherical dark matter halo enveloping the entire galaxy cluster). Of course, so far you still haven't seen exactly *how* the processes in visible matter give rise to galaxies of particular sizes and characteristics. A great deal is known about this, but scientists are far from a complete understanding. All that can be said for now is that it is a fascinating research question, and perhaps the most important goal in modern cosmology is to understand completely how the primordial pristine gas in the early Universe gave rise ultimately to all the present-day properties of galaxies.

- Older astronomy books tend to define the term 'galaxy' along the following lines. 'A galaxy is a vast assembly of luminous matter, possibly containing billions of stars, held together by the mutual gravitational attraction of its constituents.' On the basis of what you have read, propose a modified definition that is more in keeping with modern views concerning dark matter.

  A galaxy is a vast assembly of dark matter and luminous matter, possibly containing billions of stars, held together by the mutual gravitational attraction of its constituents.

There's one last caveat that has to be mentioned: how do we know for sure that dark matter is always completely collisionless? We don't. It turns out that dark matter haloes do indeed have fairly regular properties, exactly as you'd expect for collisionless dark matter. Even if dark matter is not entirely collisionless, there are strong limits on the number of collisions that go on, so any interactions between dark matter particles must be very weak and rare Perhaps dark matter is more complicated than has been assumed here, but if so, the approximation that it's collisionless looks very good indeed.

But since dark matter is transparent, how is it possible to tell where it is? And so where do these observational constraints on collisionless dark matter come from? One way is from the motion it induces in visible matter, as you have seen is the case with stars in a galaxy, and with galaxies themselves in a galaxy cluster There's another way that uses a strange but fundamental consequence of Einstein's theory of gravitation: that matter curves the very spacetime that it is in. This is discussed further in the next chapter, but for now what you need to know is that all matter, including your own body right now, is curving the space and time it's in. For everyday objects like a human body, the effect is very tiny, but for a massive cosmological object like a galaxy or a cluster of galaxies, the effect can be strong so strong, in fact, that if there is a foreground galaxy or cluster of galaxies that's close to the line of sight to some other, more distant, background galaxy, then the foreground warp of space and time visibly distorts the image of the background galaxy. Figure 3 7 shows an image of the galaxy cluster known as Abell 2218, taken using the Hubble Space Telescope. The orange galaxies are in the foreground, in the galaxy cluster itself, but you can also see a pattern of long thin arcs which are background galaxies seen through this foreground galaxy cluster. What you are seeing here is the visible effect of warped space. It's from these distortions that astronomers can figure out how the dark matter is distributed.

3.5 The Big Bang and the expanding Universe

Edwin Hubble also made another spectacular discovery that certainly ranks among the greatest discoveries ever made in science. (It's perhaps for this

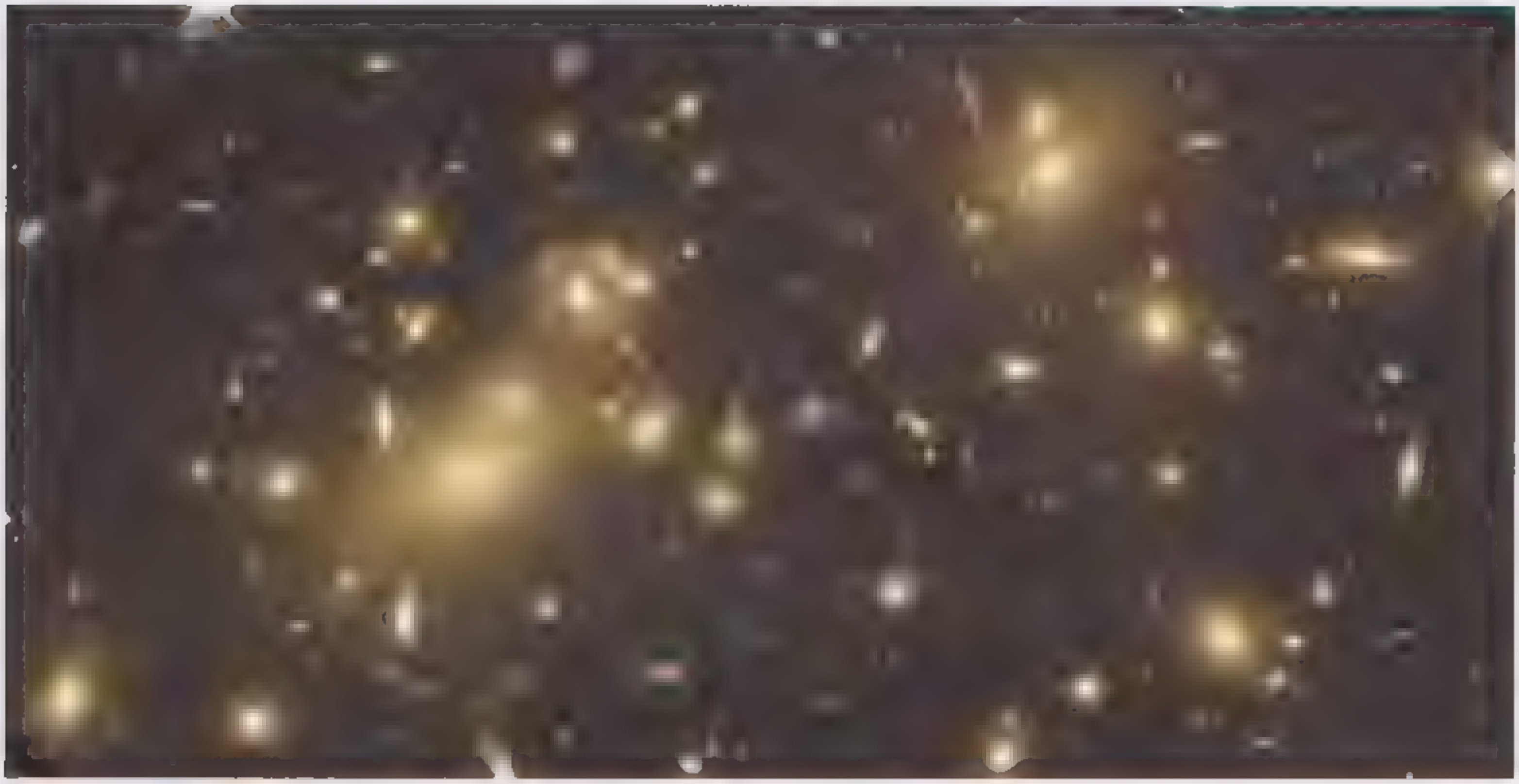


Figure 3.7 Hubble Space Telescope image of the galaxy cluster Abell 2218. The background galaxies seen through the cluster have been distorted into long thin arcs, because the foreground galaxy cluster is warping its own space

tremendous legacy that the Hubble Space Telescope was named after him, several decades after his death.) To explain it, first a little background. At the time, Einstein was struggling with formulating his general theory of relativity. This is a theory of gravity which included space and time, and it has something in common with the previous theory invented by Isaac Newton: gravity is always attractive. This means it is impossible to set up a static, eternal universe: either space must be contracting because of gravitational attraction, heading towards a **Big Crunch**, or we must be in the opposite (time-reversed) situation that space is expanding, having started in a **Big Bang**. Einstein wanted to avoid either situation, because they beg too many questions, so he introduced an additional term into his equations, called the **cosmological constant**. The effect of this cosmological constant is to bestow on space an in-built tendency to expand. This means the Universe can be set up in a perfect balance, with this tendency to expand just balancing the gravitational tendency to contract.

Edwin Hubble's spectacular idea was to take Einstein's original equations at face value. He sought evidence for the expansion or the contraction of the Universe, and discovered the Universe is expanding. It followed that the Universe started with a Big Bang. This completely changed our picture of the cosmos and our place in it, and has to rank as one of the greatest intellectual milestones in the history of human thought.

Where did the Big Bang happen? To answer this, you need to realise that it's not that galaxies are moving away from each other; space itself is expanding. When the Universe was young, space itself was smaller. At the very beginning (assuming that conclusions can sensibly be drawn that far back), space would have had zero size. If it's possible to answer the question at all, it's that the Big Bang happened everywhere, simultaneously.

If you find this difficult to imagine, it can be helpful to represent the expansion of the Universe as the surface of a balloon that's expanding. This swaps our three-dimensional space with the two-dimensional surface of the balloon. Galaxies would be things stuck on the surface of the balloon. Note that galaxies do *not* themselves expand, because they're gravitationally bound and have 'decoupled' their matter from the global expansion – you don't get any taller because of the expansion of the Universe, for example. In this picture, the expanding Universe is represented by the balloon being blown up. There is a tremendous subtlety to this picture though that is often overlooked: the distance to the centre of the balloon is a measure of *time*. When the Universe starts in a Big Bang, the balloon has a zero size and is at that central point, and as time goes on, the balloon expands away from that central point. This is why the Big Bang didn't happen at a particular point in space (i.e. at a particular spot on the surface of the balloon), but rather happened everywhere in space simultaneously.

There is a story that when Einstein heard of Hubble's discovery, he called the cosmological constant the greatest mistake of his life. In fact, the reason he said this is that it turned out his perfectly-balanced theoretical universe was unstable. Just a slight perturbation one way or another and part of it would expand or contract. It was that instability, just as much as Hubble's spectacular discovery, which led Einstein to admit this greatest mistake of his life.

In fact, as you will see in the next chapter, it turned out much later on that Einstein was partly right. The expansion of the Universe has been found to be accelerating, which has been attributed to this cosmological constant (or a generalisation of the idea known as **dark energy**). The cosmological constant is not holding the Universe together in a perpetual 'just-so', but it is contributing to driving the expansion of the Universe.

The expansion of the Universe affects the light passing through it. Imagine that two pulses of light are sent out, one second apart, from something in the distant Universe. The distance between these two light pulses is, by definition, 1 light-second. Light travels at about 300 million metres per second, so the distance between these pulses is about 300 million metres. As these two pulses travel across the Universe, space is expanding, so that distance of 300 million metres becomes larger and larger. By the time the light pulses are received by a telescope on Earth, the distance between the pulses would be much larger than the original 300 million metres. Light still travels at 300 million metres per second though, so the way this is experienced on Earth is that the difference in the arrival times of the two pulses is much longer than 1 second.

This means that if someone used telescopes to watch a clock in the distant Universe, the clock would seem to be running slowly, in the sense that the second hand would seem to be moving too slowly. It would take more than a minute for the second hand on this distant clock to rotate once. Now, there aren't any handy cosmological clock faces to watch, but there's the next best thing: some types of supernovae brighten and slowly become fainter with very characteristic rates. Looking further and further into the distant Universe, the supernovae seem to brighten and become fainter more and more slowly. This is known as **cosmological time dilation** (time dilation means that someone else's time, as seen from here, seems to be running slowly).

- Suppose in that distant part of the Universe, someone is watching *us*. Would they think *our* clocks are running slowly or more quickly?

  They would also think our clocks are running slowly, for the same reasons we see theirs as slow.

The expansion of the Universe has another effect on the light passing through it. This time, instead of imagining two pulses of light, imagine following two particular peaks of a light wave as it travels through the expanding Universe. (Look back to Section 2.2.1 now if you are unsure.) Just as the expansion stretched the distance between the two pulses of light, the expansion will also stretch the distance between the two peaks of the wave. By the time the wave is received on the Earth, the wavelength of the light (i.e. the distance between two successive peaks of the wave) will be much longer than when the light was emitted. Now, a longer wavelength means the light will appear redder This is shown in Figure 3.8.

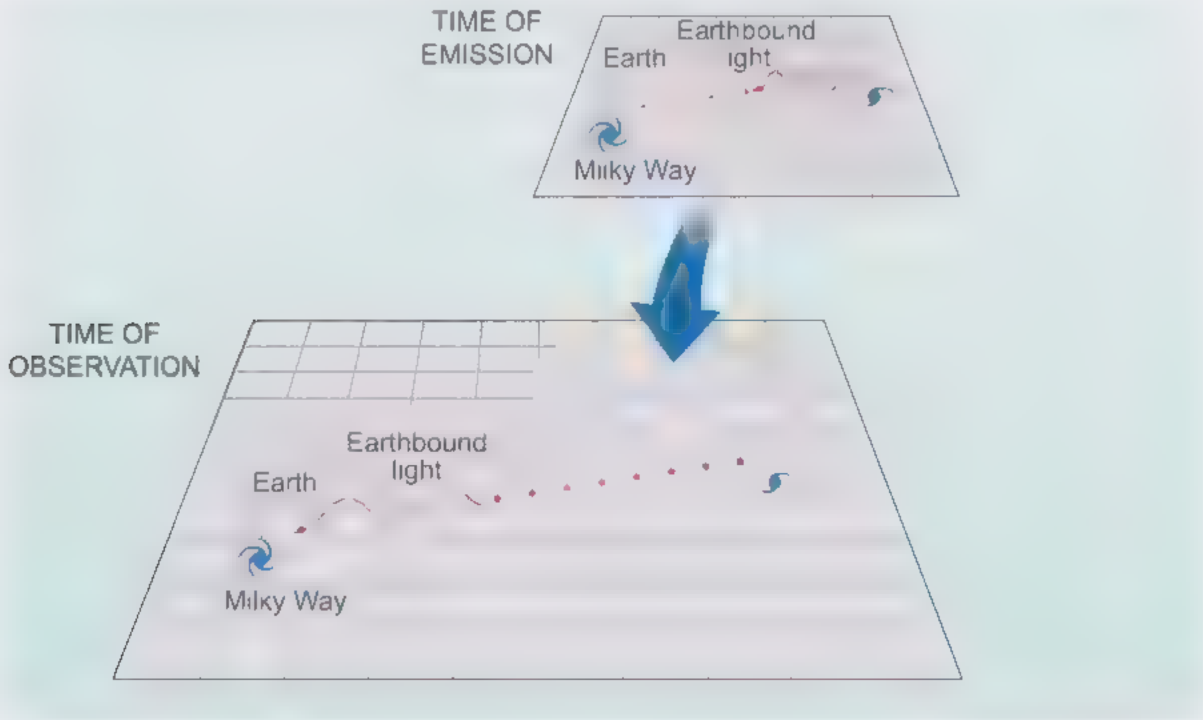


Figure 3.8 Expanding space causes redshift and Hubble's law.

This means that when our telescopes observe distant parts of the Universe, the light from those distant objects is shifted to the red (i.e. the wavelength becomes longer). This is known as **redshift**. The further away a galaxy is, the longer the light takes to get to us, and so the more stretching the light wave has had in the meantime. So, the bigger the redshift, the further away the galaxy is. Also, the bigger the redshift, the longer the light has taken to get to us, so the further back in time we're seeing.

The way that Edwin Hubble discovered the expansion of the Universe is that he found that galaxies that are more distant have higher redshifts. This relationship became known as **Hubble's law**. To demonstrate this, he needed a way of determining distances other than using redshift. He used Cepheid variable stars, which you have already met in Section 2.5.4. These stars have known luminosities. By comparing this known luminosity with how bright they appear, he could deduce the distance to the galaxy, which he could then compare to the redshift.

Now, there is an alternative explanation of cosmological redshift that you will find in many textbooks, but which is mostly wrong: the Doppler effect, which you met in Section 2.5.3. An example is the change in pitch of an ambulance siren as it passes you – as it approaches you the notes are higher, and as the ambulance speeds into the distance away from you the notes are lower (Figure 2.15). Lower notes mean longer wavelengths. Since the expansion of the Universe means that more and more distant things are moving faster and faster away from you, sometimes cosmological redshift is attributed to this Doppler effect. However, things moving in unexpanding space are a very different physical situation from having space itself stretch. As you have seen, the Universe is a dynamic place, and as galaxies move relative to each other, there are indeed Doppler redshifts (as well as Doppler blueshifts for approaching galaxies). However, once you get far enough out, these Doppler shifts become small compared to the cosmological redshift caused by the stretching of space.

■ Does the Andromeda galaxy appear blueshifted or redshifted to us?

Our galaxy is falling towards Andromeda, so it appears blueshifted. Note that this is a genuine Doppler shift in this case, not a cosmological redshift.

[illegible] The large-scale structure of the Universe

Let's resume our tour of the Universe by moving outward to the next largest scale. The next nearest galaxy cluster is the Coma cluster, containing over 1000 galaxies. Both the Virgo and Coma clusters are part of a 'sheet' or 'wall' of galaxies and clusters known as the **local supercluster**. This is not a bound structure in the same way as individual galaxy clusters, but still exists because of the gravitational attraction of its components. Moving further and further out the evidence of voids emptying out and collecting in sheets and filaments can be seen. This is sometimes described as the **cosmic web**. Figure 3.9 shows the structure of the Universe seen in two strips on the sky, measured using the Anglo-Australian Telescope. The number of galaxies thins out towards the edge, because only the brightest galaxies can be seen at that distance – it's not because there are fewer galaxies out there.

To look deeper into the Universe, astronomers can only observe smaller patches of sky. This is because telescopes can spend their time either

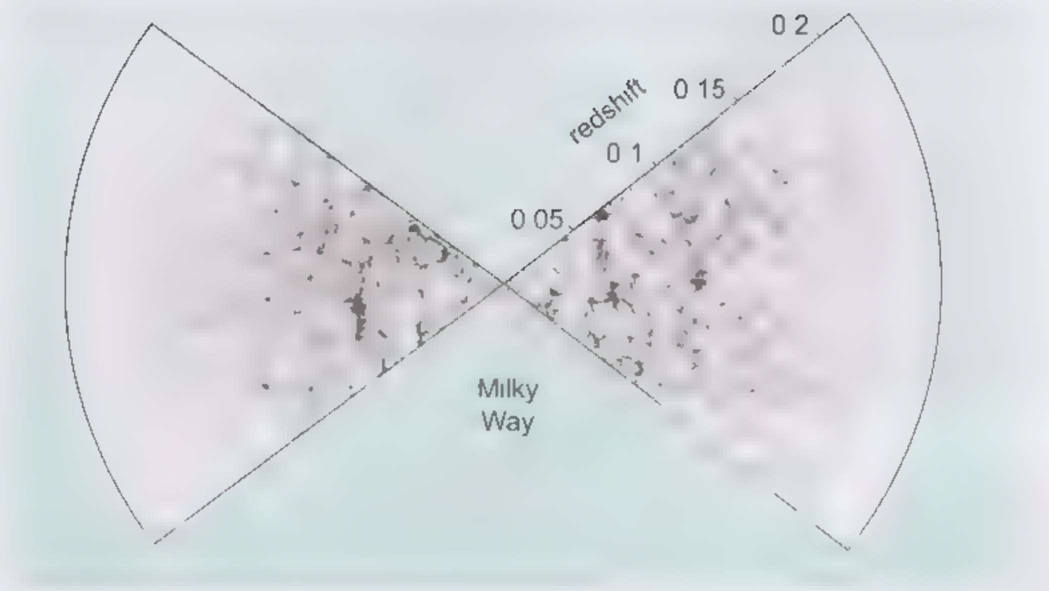


Figure 3.9 The large-scale structure of the Universe as measured by the galaxies detected by the Anglo–Australian Telescope.

sweeping the sky quickly in a shallow survey, or scanning slowly to make a deep survey. The most famous deep surveys, and some of the most spectacular, have been the ones made by the Hubble Space Telescope, or HST as it is known to professional astronomers. In 1996, HST invested ten consecutive days staring almost entirely at one small patch of sky in the constellation of Ursa Major, shown in Figure 3.10. This was known at the time as the Hubble Deep Field, but after a second similar observation was taken in the southern sky, it became known as the **Hubble Deep Field North**. Because of the extraordinary depth of this deep-sky image, the light from its most distant galaxies took most of the history of the Universe to reach us. In other words, it shows those galaxies as they were early on in the history of the Universe. The further back we look, the longer the light has taken to reach us, so the further into the history of the Universe we are probing. Because the Hubble Deep Field North covers only a small patch of sky, it is surveying a very long and thin volume of the Universe.

The HST is above the Earth's turbulent atmosphere, which makes its images much sharper than ground-based telescopes can achieve. It was this, as much as the depth, that made the Hubble Deep Field North so successful. One discovery was that the Hubble tuning fork of galaxies breaks down in the early Universe. It's an extraordinary and astonishing thing that it's possible to see the stages of the assembly of spiral discs throughout almost the entire history of the Universe.

There is also the Hubble Deep Field South, and other deep Hubble Space Telescope surveys. The numbers and types of galaxies are remarkably similar along these different lines of sight, implying the Universe looks uniform on the largest scales.

There is another approach to measuring the largest-scale structures in the Universe, and that is to use **quasi-stellar objects**, often abbreviated **QSOs** or referred to as **quasars**. The name 'quasi-stellar' refers to the fact that they

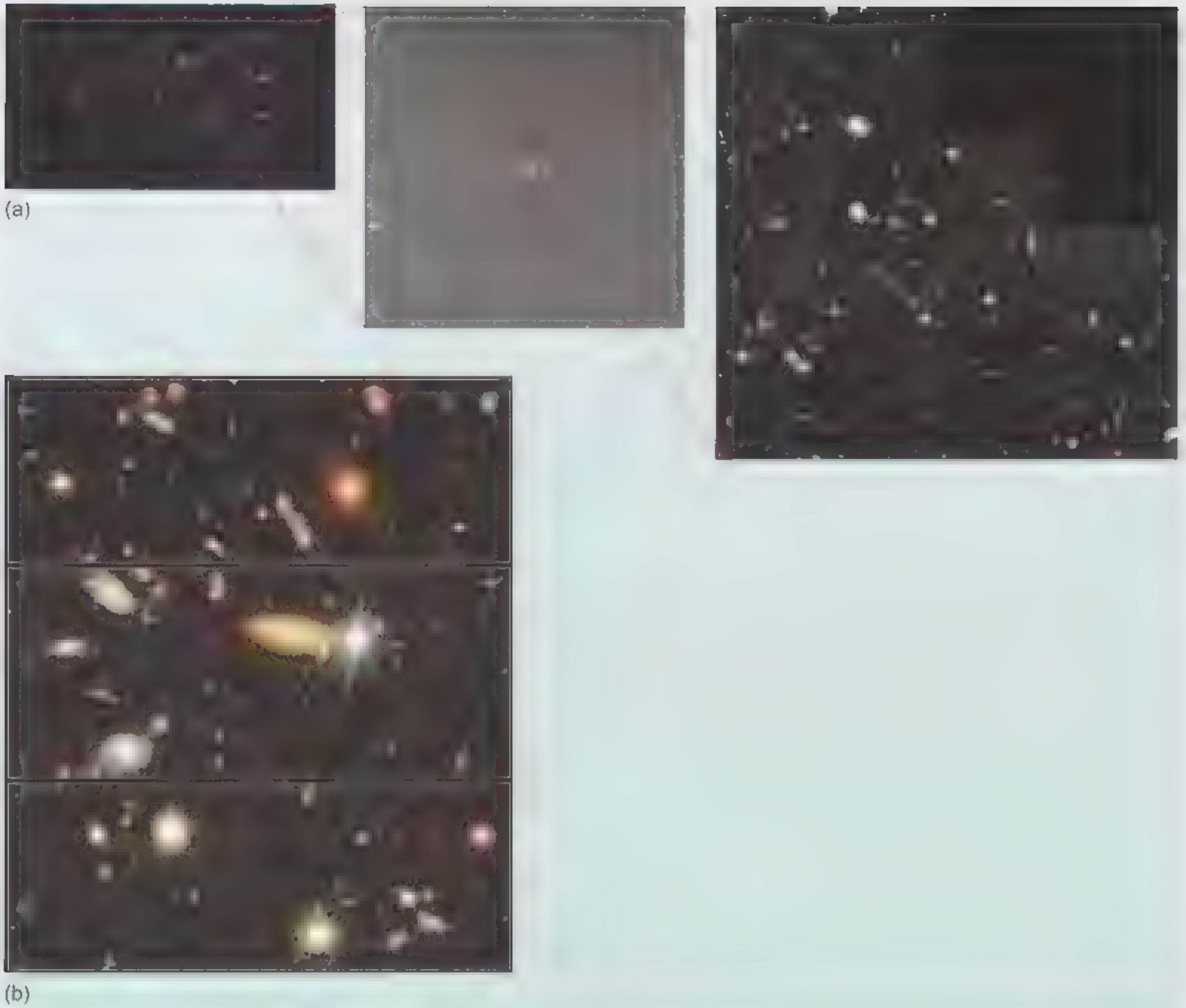


Figure 3.10 (a) The Ursa Major constellation, a zoom of an unremarkable part of it, and a further zoom which is the location of the Hubble Deep Field North. (b) Details from the Hubble Deep Field North. Some galaxies are spirals and ellipticals similar to those seen in the local Universe, but there are also strange blue blobs and chains (such as in the middle of the top panel) that have no counterparts in the present-day Universe.

looked like stars in early optical images in the 1950s, but when it was discovered that they had cosmological redshifts, it was clear that they are not stars but astonishingly luminous and distant point-like objects, often much more luminous than entire galaxies! The physical sizes must be very small, because the light from quasars often varies strongly over the course of weeks, meaning their physical sizes must be less than light-weeks (one light-week is about 1 52 of a light-year, or about 1200 times the distance from the Earth to the Sun) What could cause these enormous luminosities from something so compact? The consensus view is that these are **supermassive black holes** in the centres of galaxies, where 'supermassive' means a million or even a billion times more massive than our Sun. Galaxies that contain quasars are known as **active galaxies**, and the black holes active in their centres are known as **active galactic nuclei**. There is even a supermassive black hole in the centre of the Milky Way, though it's not currently an active galactic nucleus.

You will meet black holes in Chapter 6, but for now what you need to know is that once matter has passed inside a black hole, it's completely lost to the rest of the Universe. Nevertheless, the process of falling towards the black hole releases a lot of energy. A sparse amount of matter falling in would end up moving extremely quickly by the time it reached the black hole, but there's another possibility if there is a great deal of matter falling in: the matter could form an **accretion disc**, which becomes very hot (approaching 1 million degrees K) and which consequently radiates away energy as thermal radiation. It's this radiation that results in quasars being so luminous. It's a curious paradox that the blackest things in the Universe, black holes, can be surrounded by some of the most luminous matter in the Universe.

Because of these tremendous luminosities, quasars can be seen more easily to great distances than regular galaxies. Figure 3.11 shows the largest-scale structure of the Universe as demonstrated by the quasars discovered by the Anglo Australian Telescope. Just as in Figure 3.9, only the brightest objects can be seen at the greatest distances, which is why there appears to be a decline in the numbers of quasars at the greatest distances. This time though, there's also a deficit at the centre, which is at low redshifts (i.e. close to us). This is because quasars are much rarer today than earlier in the history of the Universe (unless of course the Earth is exactly in the centre of a hole in the large-scale distribution of quasars with perfectly symmetrically graduated edges). It's not entirely understood why quasars have become rarer, but it's probably connected with the formation and evolution of galaxies, because the masses of supermassive black holes turn out to be very closely related to the properties of the galaxies they are in.

In any case, Figures 3.9 and 3.11 make it clear that on the very largest scales the Universe appears **homogeneous** and **isotropic**. Homogeneous means it appears the same everywhere, and isotropic means it has no particular cosmic direction. If the Universe had a constant uniform magnetic field, the direction of that field would make the Universe *anisotropic*, meaning 'not isotropic'. If there were more galaxies seen from the Northern Hemisphere than the Southern, the Universe would be *inhomogeneous*, meaning 'not homogeneous'. (In fact in the nearby Universe there is a slight difference seen from the North and the South, but at sufficiently large distances these differences disappear.)

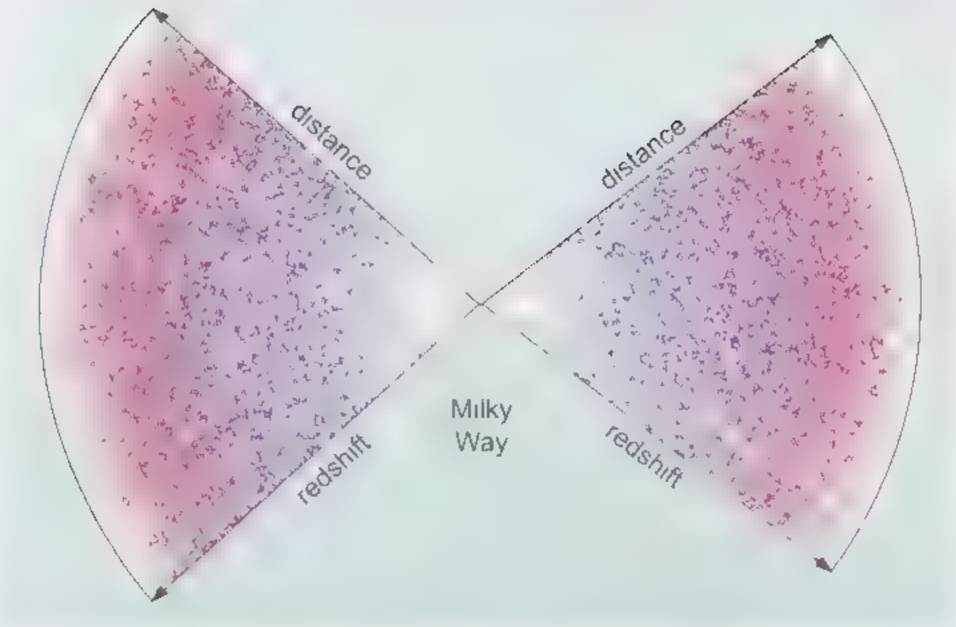


Figure 3.11 The large-scale structure of the Universe (even larger than Figure 3.9), as traced by quasars detected by the Anglo Australian Telescope in two strips on the sky. The diagram goes out to a redshift of about 3 corresponding to a distance of about 6.5 billion parsecs.

3.7 Approaching the edge

This final section of the chapter looks almost to the edge of the observable Universe, with the most distant objects that have ever been found, and in the next chapter you will peer even further to the farthest thing that has ever been seen.

You will be familiar with rainbows, where the Sun's light is split up into its constituent wavelengths, with red (longer) wavelengths at one end and blue (shorter) wavelengths at the other. As you saw in Chapter 2, astronomers generalise this idea with a plot of a spectrum. An example is shown in Figure 3.12. This is a spectrum of a quasar known as 3C 273. The horizontal axis is the wavelength λ (blue to the left, red to the right), and the vertical axis is how much light is being received at each wavelength. The spectrum is fairly flat, except for a few bumps that have been labelled. These bumps are known as **emission lines**, and are caused by light from particular gases (Figure 2.2). Their characteristic wavelengths are caused by the individual atomic structures of the gases. Characteristic features like these make it possible for astronomers to measure redshifts: by knowing the wavelengths of these features from Earth-based laboratory measurements of these gases, then comparing the wavelengths that they're seen at in the cosmological object, one can calculate how far the lines have shifted to the red.

Calculating redshifts

Astronomers calculate the redshift by taking the change in wavelength of an emission line, then dividing it by that emission line's original wavelength. It turns out that this trick of dividing by the original wavelength gives you an answer that doesn't depend on which emission line you choose, so any emission line will give you the same value for the redshift for that galaxy. For example, if a galaxy has an emission line with an original wavelength 500 nanometres (nm) that is then redshifted to 1000 nm, the galaxy is said to be at a redshift of (1000 − 500)/500 = 500/500 = 1.

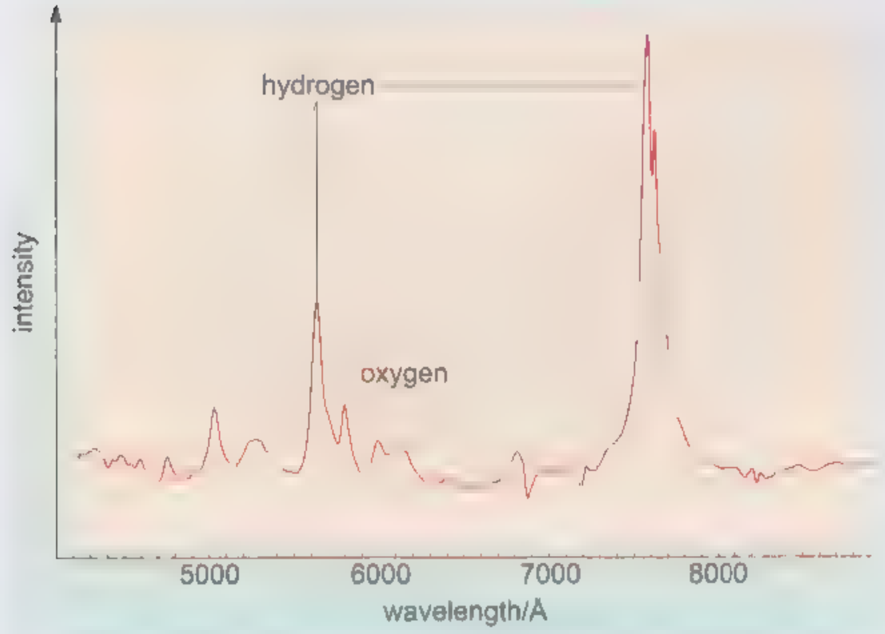


Figure 3.12 The spectrum of the optical light from quasar 3C 273. The wavelength axis is in ångströms, where 1 ångström = 10^{10} metres. The vertical axis is the intensity of light that's being received at each wavelength by telescopes.

- What is the name of the light that has wavelengths that are just too short to be seen by the human eye?

  The answer is ultraviolet.

- What is the name of the light that has wavelengths that are just too long to be seen by the human eye?

  The answer is infrared.

For the most distant cosmological objects, all their optical and ultraviolet photons are shifted into the infrared.

What colour is the sky? It's blue in daytime, black at night. What about in space? Black is how it would appear, but if you had extremely sensitive eyes (or a very expensive telescope) you'd be able to see all the light from all the stars and all the galaxies that have ever existed. This would probably come out as a dull reddish colour. Now let's imagine it's possible to plot the spectrum of this background light, and let's extend this into infrared light. When astronomers make this measurement (Figure 3.13), there's a surprise:

there's a bump at optical wavelengths that extends to longer wavelengths (known as the near-infrared), then there's a minimum at a wavelength of about 15 microns. At longer wavelengths still, there's a second bump that can be attributed to thermal infrared radiation from heated dust. There's as much energy in this second (infrared) bump as there is in the optical bump. This immediately tells us something profound about the Universe: roughly speaking, for every two optical photons (recall from Section 2.2.1 that a photon is a particle of light) that have ever been emitted by *any* star or *any* galaxy or *any* accreting black hole, *ever in the history of the Universe*, one of those photons has been absorbed by dust (probably quite soon after it was emitted), and its energy re-radiated as thermal radiation by that dust. And it's possible to tell this just from the colour of the sky in space.

Note: the far infrared background light is *not* redshifted light from the optical background light. It is thermal emission from heated dust that has absorbed photons of visible light.

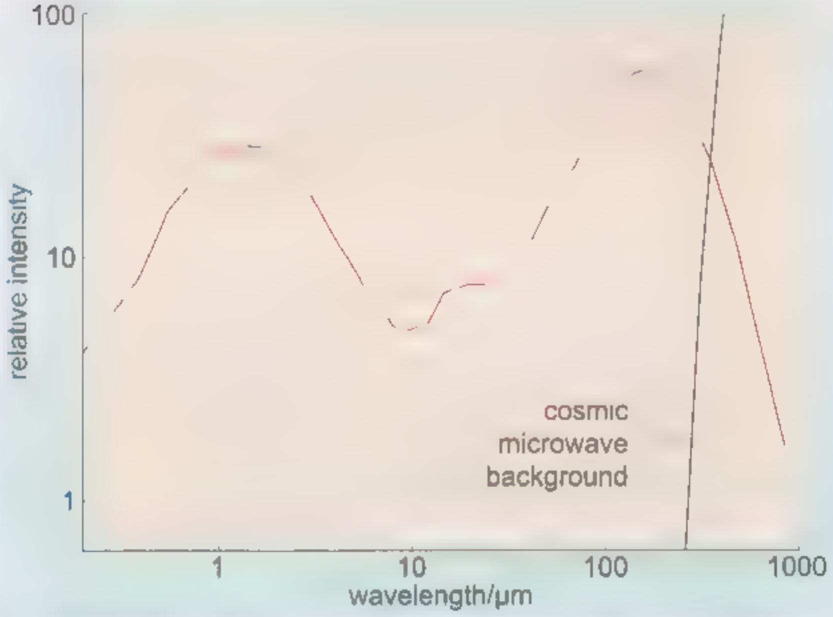


Figure 3.13 The spectrum of the extragalactic background light. The wavelength axis is in microns. The almost vertical line on the right-hand side is the cosmic microwave background, which you will meet again in Chapter 4. Its peak is off the top of the figure.

This is one of the reasons that astronomers have launched many infrared space telescopes that complement the optical and near-infrared work of the HST. The sky can look very different in infrared light. To give you an example, Figure 3.14 shows the constellation Orion, as seen by the AKARI space telescope at a wavelength of 140 microns. Unlike our optical view of the night sky, the stars are not what dominate the image. Instead, what's seen is infrared light from dust in the Galaxy The Orion Nebula, just below Orion's belt, is a bright knot of warm dust, much of which just looks dark at optical wavelengths. The Orion Nebula is luminous in this infrared light because of star formation happening inside its opaque dust cloud. (There's a slight subtlety: if our eyes detected light at 140 microns, we wouldn't see exactly what is in Figure 3.14 because our atmosphere is quite opaque at that wavelength, so this could only be seen from high-altitude balloons or aircraft or space telescopes.)

- In what ways would the night sky look different if we lived in an elliptical galaxy (where very few new stars are forming)?

  If we lived in such a galaxy we would still see stars, but there would be no galactic disc and so no band of light comparable to the Milky Way, and no nebulae comparable to the Orion nebula where stars are forming.

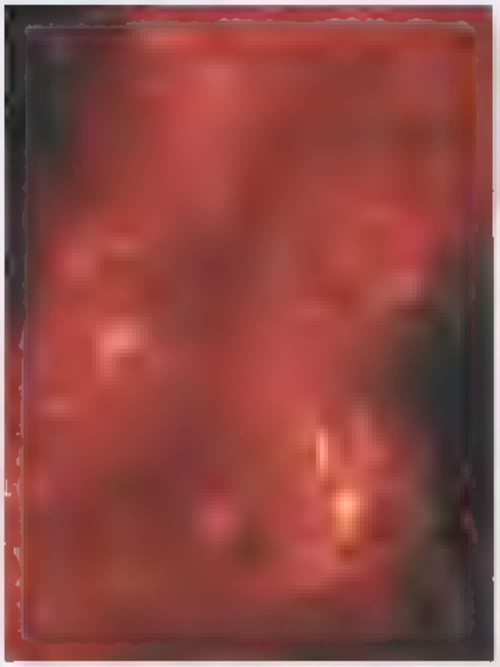


Figure 3.14 The constellation of Orion, seen in the infrared.

Space observatories such as the HST and infrared space telescopes have been used to find out information about how galaxies formed and evolved. If we are interested in where their stars came from, this means finding out where and when the stars formed. With star formation there are two particularly relevant processes for us in this section:

(a) there are very bright young massive stars that generate a great deal of optical and ultraviolet light (these will be discussed in more depth in Chapter 5) and

(b) a great deal of dust can be generated which can obscure our optical view.

Therefore, to find out how galaxies formed and evolved astronomers need to look for galaxies forming stars with both optical and infrared telescopes.

Some of the greatest technical achievements of these space telescopes, including (but not limited to) the HST, have been the measurements of the cosmic history of star formation. This is the number of stars forming per year, per unit volume of the Universe (this is normally adjusted to account for the expansion factor). Now, you might reasonably ask why you should care about this history. One reason you might care is that everything you see around you – every table, chair, wall, person, rock – in fact, pretty much *all* visible matter in the Universe that isn't hydrogen or helium – *all* of this has been generated in nuclear reactions in stars. These reactions are discussed later in this book but for now the important point is that the formation of stars is telling you about the creation of 'stuff' in the Universe. This is why cosmologists regard this cosmic star formation history as so fundamental.

Figure 3.15 shows a recent compilation of measurements of this cosmic star formation history. As we look back in time out to redshifts of between 2 and 3, we see that there was a lot more star formation going on than there is today. Looking even further back, we see that the star formation rate declines again. This means we've found the peak epoch for the creation of heavy elements in the history of the Universe, and it's these heavy elements that make up planet Earth and everything on it.

To finish off this chapter, we'll tell you about a couple of ongoing scientific mysteries related to this cosmic history of star formation. A curious follow-up discovery has been that extremely luminous star-forming galaxies contributed a large proportion of the cosmic star formation during that peak, while today they contribute a very small proportion. Why were huge violent star formation episodes such big contributors then, but not now? This is still not well understood.

Figure 3.15 has the highest-redshift data that is available at the moment (2012) Even to the HST, the most distant objects are un-photogenic, unprepossessing blobs. Nevertheless, it's clear that the cosmic star formation rate in Figure 3.15 goes down quickly for the highest redshifts. This means there's much less light from young stars. However, we know from the state of the gas between the galaxies at that point that there must have been a lot more ultraviolet light around at the time, and that this ultraviolet light must have been emitted even earlier. So, what generated this first light from the first objects in the Universe? Was it young galaxies violently forming stars, or was it quasars with black holes accreting lots of hot luminous gas? No-one knows for sure at the moment, but new telescopes might give the answer. HST's successor, the James Webb Space Telescope, will specialise in infrared light and might directly detect clumps of the very stars or black holes that made the first light from any objects in the Universe, and ended the dark age of the Universe.

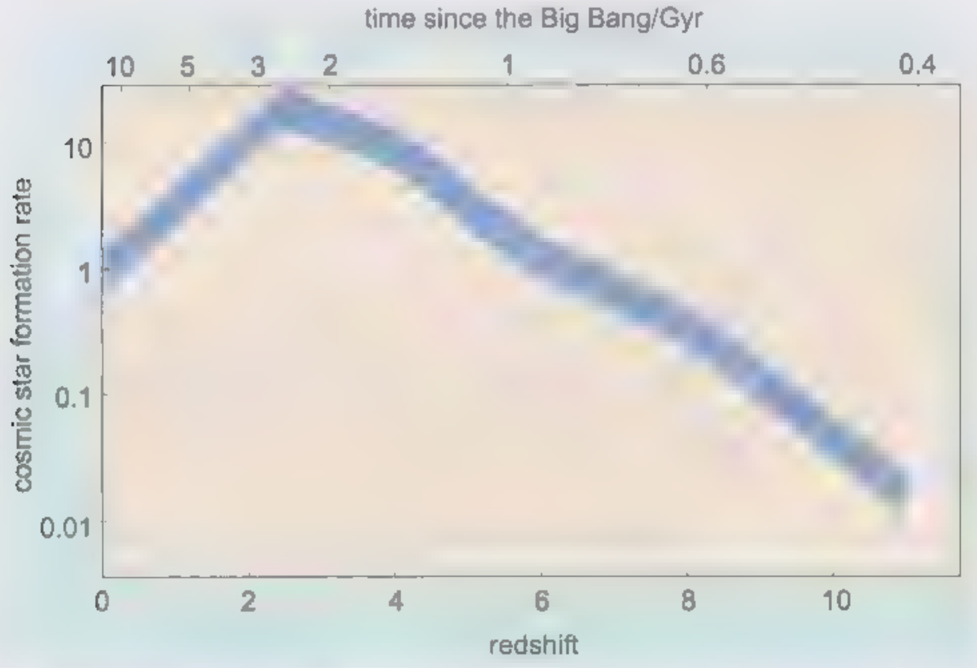


Figure 3.15 The cosmic star formation history The lower horizontal axis shows the redshift, while the upper horizontal axis marks the age of the Universe in billions of years at those redshifts. Note that these ages aren't evenly spaced. For the time axis, 1 Gyr = 1 gigayear = 10^9 years. For the vertical axis, a cosmic star formation rate of 1 is equivalent to 1 Sun-like star forming per century in a volume of 1 cubic megaparsec.

Activity 3.1 Why do astronomers use non-optical telescopes?

The estimated time for this activity is 3 [illegible]

In this activity you will cross-fade two images from a galaxy survey that were taken at different wavelengths to look for star-forming galaxies. The detailed notes for this activity are in the 'Activities' section of the module website.

End-of-chapter questions

Question 3.1 If you had a glass full of dark matter, what would it look like, and what would happen to the dark matter?

Question 3.2 Figure 3.16 shows three galaxies. On the basis of its appearance, classify each one as a spiral, elliptical or irregular.

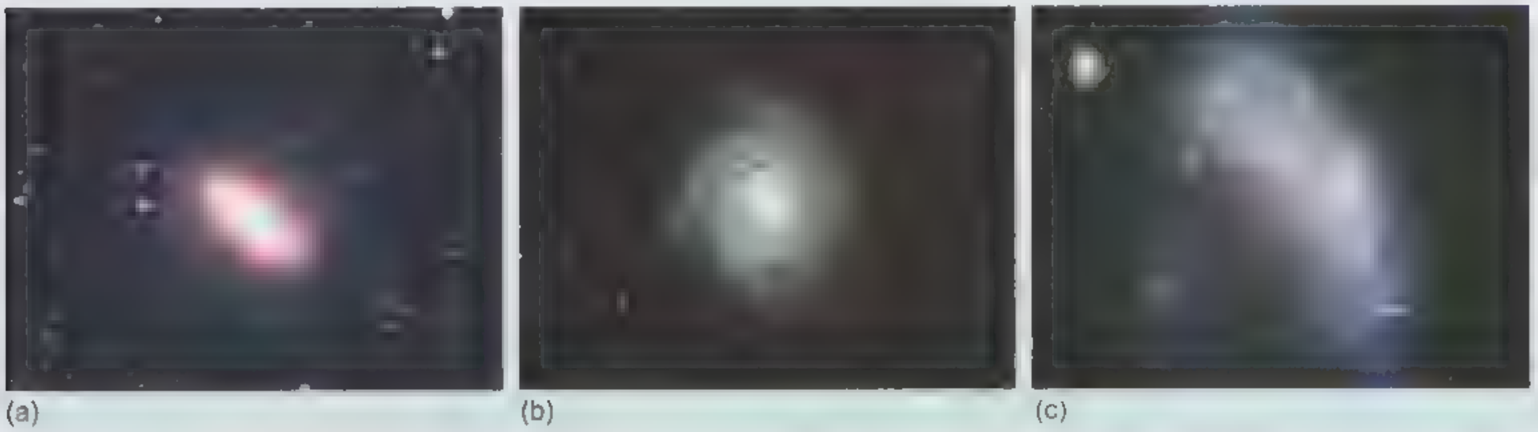


Figure 3.16 Photographs of galaxies for Question 3.2

Question 3.3 Describe an arrangement of galaxies in the Universe that would be homogeneous but anisotropic.

Question 3.4 Figure 3.17 (overleaf) shows a way of picturing the expanding Universe. You draw galaxies on an elastic band, then stretch the band to show the expansion of space. However, there's something subtly misleading about this picture. What is it?

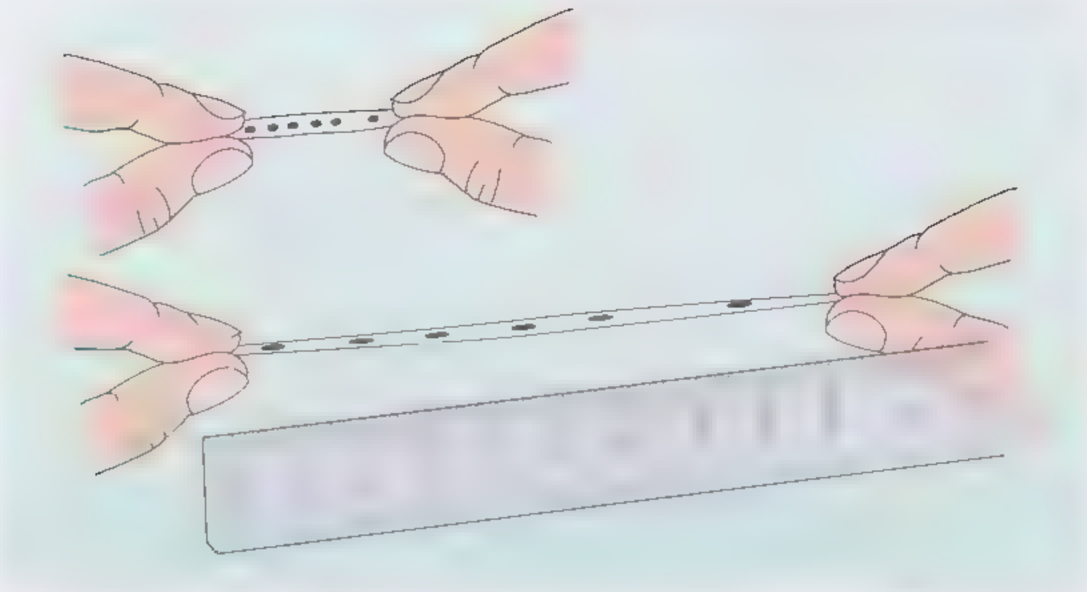


Figure 3.17 Modelling galaxies in a uniformly expanding Universe.

Question 3.5 At the time of writing, the most distant quasar ever discovered has an emission line observed in the infrared (Figure 3.18). This emission line is called Ly α or Lyman α (pronounced 'Lie-man alpha') and has an original ultraviolet wavelength of 121.6 nm. Estimate the observed wavelength of this line and calculate the redshift of this quasar.

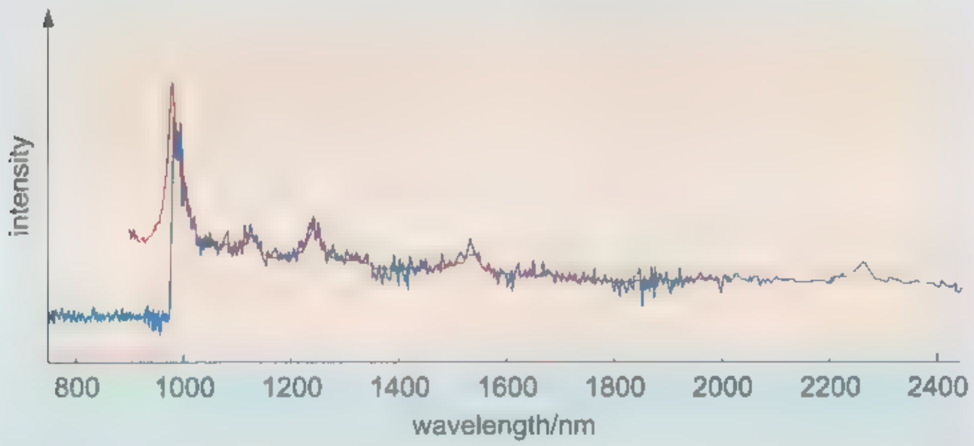


Figure 3.18 The spectrum of the most distant quasar seen so far. The red line is an average spectrum of less-distant quasars, shifted artificially for comparison. The largest peak is the Lyman α emission line. The sharp cut-off below the Lyman α line is caused by intervening absorbing gas.

Now go to the module website and do the remaining activities associated with Chapter 3.

Activity 3 2 - 60mins
Activity 3 3 - 45 mins
Activity 3.4 - 15 mins
Activity 3 5 - 15 mins

Chapter 4 The edge of the Universe

4.1 Introduction

What is the most distant thing that can be seen with any telescope? This is currently the cosmic microwave background (CMB, Figure 3.13). This is the light from the time when the Universe was so small and dense, it was opaque. In a sense, it represents the edge of our Universe, or at least that bit of the Universe that can be seen with telescopes. This chapter will touch on some of the deep questions about the origin of the Universe that are posed by this very distant light.

Developing your writing skills

All scientists need the ability to communicate clearly, especially through their writing. Developing good writing skills, particularly when they don't come naturally, takes considerable time and effort but it is essential to develop them. This module alone cannot help you achieve this, but you can use the study materials to improve your writing skills. Much of this chapter is descriptive; it deals in a simple way with matters that are very complicated and far-reaching. Many of its assertions are accompanied by phrases that limit the circumstances to which they apply, or even indicate that there are doubts about their validity. As you read this chapter, look out for these phrases. Ask yourself what is really being said and whether it could be said more briefly. Would a briefer statement mean a vital condition or reservation is omitted? Would a briefer statement be clearer?

At the end of every few paragraphs, possibly at the end of every paragraph, stop and ask yourself whether you cou.d summarise what you have just read. If not, try reading it again. Also remember that, important as writing is, it is not the only way of communicating scientific information. Always watch out for places where information might be better presented as a picture, table or chart rather than a block of prose. Part of the skill of good writing is knowing when to stop.

You will be given some chances to practise these skills in the End-of-chapter questions, but you should consciously try to develop them in everything you do.

4.2 The surface of last scattering

Imagine you can rewind the history of the Universe, so it becomes smaller and denser, and smaller and denser. You wind the history of the Universe back to before there were any stars or galaxies or black holes, and the Universe is therefore dark. These are the dark ages of the Universe.

You can keep rewinding the history of the Universe, making it still smaller and still denser, until you reach the point where it's so small and dense that it's opaque, like the surface of the Sun. This is long before the time of stars though. If you're floating somewhere inside this early Universe, everywhere would look like an almost perfectly uniform glowing fog. This fog would be glowing because all the gas would be hot, at a temperature of about 3000 K (for comparison, the surface of the Sun has a temperature of roughly 5000 K). By rewinding the history of the Universe, you've compressed the gas in it. In general, compressing a gas heats it up, just as expanding a gas generally cools it down.

- Compressing the gas in the Universe means you'd be looking through more material. To see why looking through more material might make it appear opaque, imagine looking through a mile-wide stack of panes of clear glass. What would you see?

  You wouldn't see much, if anything. Almost no light would make it from one end to the other. You might see scattered light that originally came in from the sides, depending on how big the panes of glass were, and you might see light from your side that's scattered back towards you. It turns out that blue light is scattered most easily (that's why the sky is blue), and glass has a slight green tint, so looking into this mile-wide stack of glass would probably be like gazing into a very deep green–blue sea, even though individual panes of glass look almost perfectly transparent.

Now imagine winding the history of the Universe forward again. You reach the time when the Universe becomes transparent, and this fog would clear around you. But you wouldn't see the fog clear everywhere simultaneously, because it takes time for the light from more distant parts of the Universe to get to you. What you would see is a receding bank of fog, as the light from more and more distant parts finally reaches you.

This receding bank of fog is called the 'surface of last scattering'. To see why, imagine this instead from a photon's point of view. When the Universe is opaque, the photon is being scattered often. Once the Universe becomes transparent, these scatterings quickly become very rare. The final scattering of the photon would appear to you to have been at the surface of that receding bank of fog.

We are still in exactly this Universe. What has been referred to as the receding bank of fog is exactly what astronomers are observing when they refer to the **cosmic microwave background**. It's clearly a cosmic background, but why is it called a *microwave* background? To see why, let's go back to the last moment when the Universe was opaque. The temperature was around 3000 K, so the gas was glowing. The colour of that glow would be a bit redder than the present-day Sun, because it's slightly cooler than the surface of our present-day Sun. It would look a bit like the colour of the Sun less than an hour after sunrise. The midday Sun's light peaks around the yellow part of the spectrum, but this light would peak in the red at a wavelength of about 1 micron, only just beyond the limit of human red perception (about three-quarters of a micron). This receding bank of fog would therefore have a red tint. Now, as time goes on, the Universe expands, and the light from this receding bank of fog would begin to be stretched. This would make the wavelengths longer, shifting the light even further to the red. This stretching has now reached a factor of about 1000, so the intensity of light that originally peaked at around 1 micron now peaks at around 1000 microns, or 1 millimetre. Wavelengths of light longer than about a millimetre are known as microwaves. The light from this receding bank of fog now straddles the border of the microwave range, but it's nevertheless been given the name cosmic *microwave* background.

4.3 The observable Universe and the horizon problem

So, we are surrounded by this (apparent) receding bank of fog, that's giving off light that has been redshifted into microwave wavelengths. If you observe it with a telescope sensitive to microwaves, what does this redshifted bank of fog look like? There is one thing that's very striking: it's very, *very*, uniform. You can buy globes that show the positions of stars on the sky, but if you wanted to paint a globe to show what this receding bank of fog looks like on the sky, you'd have trouble making the paint uniform enough. The deviations from perfect smoothness are about one-thousandth of the brightness of the cosmic microwave background, and most of that is just a subtle feature caused by the Doppler effect of the Earth's local motion. If you correct for that, you find that the cosmic microwave background is almost exactly the same everywhere, but there are nonetheless very subtle irregularities. It turns out that the observed cosmic microwave background has a spectrum that's the same as an object at a temperature of 2.73 K, but the irregularities show up as temperature fluctuations of a few tens of microkelvin (one microkelvin, or μK, is a millionth of a K), so the irregularities are very subtle indeed. Figure 4.1 shows these tiny irregularities in the cosmic microwave background, but remember that if you wanted to see the cosmic microwave background itself instead of the ripples on it, it would just be one big block of colour.

This background is so smooth that the extreme smoothness needs some explanation. You might ask, 'why shouldn't it be smooth?' but the following argument might convince you otherwise.

The fastest that any signal can travel in the Universe is the speed of light. This means that the furthest a signal could travel in one year is the distance light travels in one year, also known as one light-year. The speed of light is about 300 million metres per second, and there are about 30 million seconds in a year, so one light-year in metres is about 300 million × 30 million, which is about 9000 million million metres. In scientific notation this is 9×10^{15} metres. In two years, the furthest a signal can travel is two light-years, or $2 \times 9 \times 10^{15}$ metres, or 18×10^{15} metres.

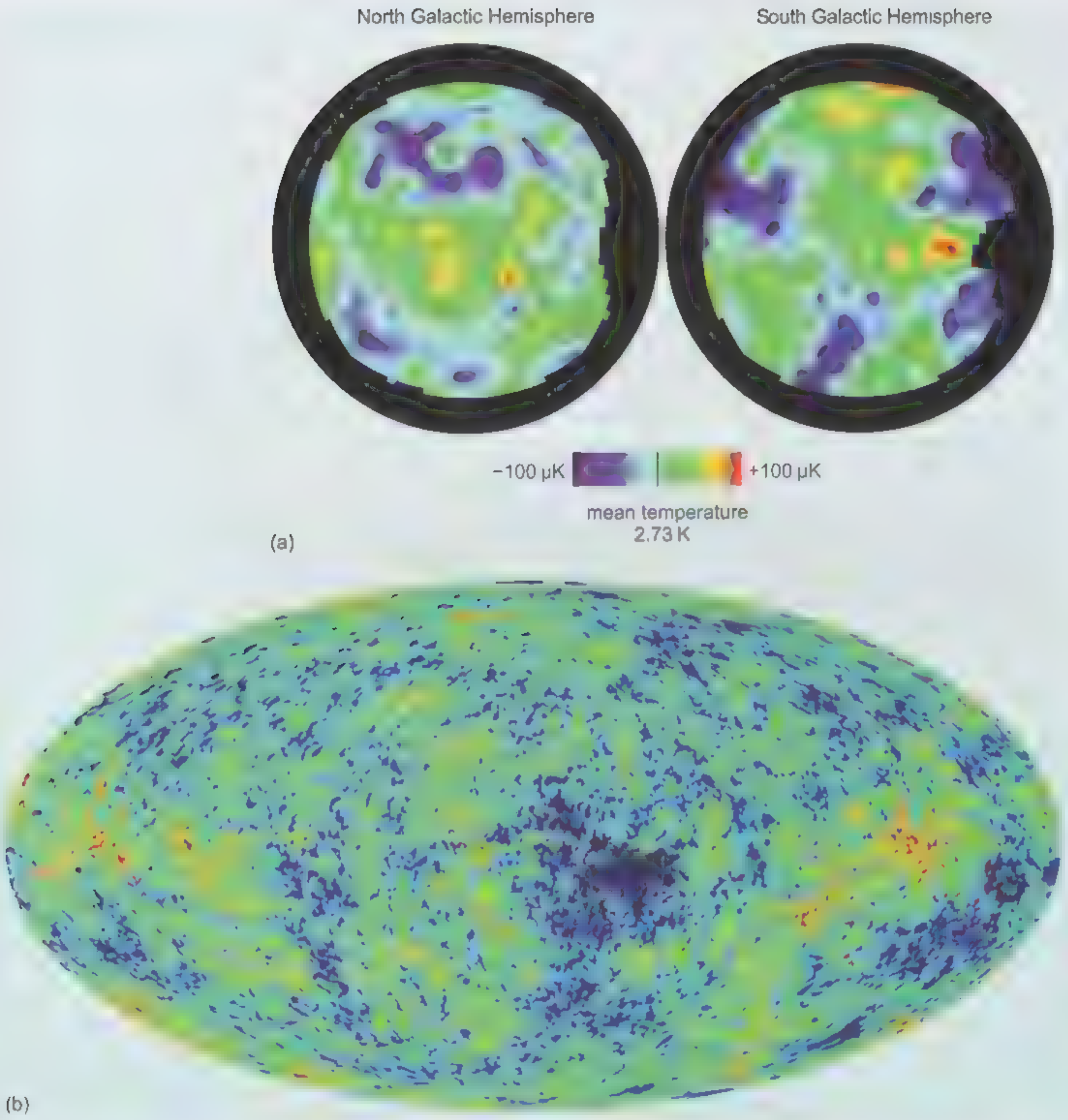


Figure 4.1 (a) This colour-coded map shows departures from uniformity in the cosmic microwave background radiation over the whole sky. The two panels correspond to two 'halves' of the sky, projected onto a flat picture. The scale represents the temperature either side of the average temperature of 2.73 K. Violet regions are slightly cooler than the average value of 2.73 K; red regions are slightly hotter, by about 100 microkelvin (μK). Most of the variations seen are believed to represent localised variations in the density of matter at a time 300 000 years after the Big Bang when this radiation interacted with matter for the last time. This map is the final result after 4 years of operation of the Cosmic Background Explorer (COBE) satellite. (b) This map, produced by the Wilkinson Microwave Anisotropy Probe (WMAP) satellite, was released in 2003. It shows the whole sky. Ripples in the temperature of the microwave background are seen here on much finer scales than was possible with COBE.

If you had wanted someone in a distant galaxy to receive your signal today, you'd have had to send the signal much earlier. The more distant the receiver, the earlier you'd have had to send your signal. But there's a limit to how early you can go, because eventually you approach the Big Bang. The age of the Universe is about 13.7 billion years, which sets a limit to how far away you could have sent a signal. You might think that the maximum distance your signal could reach is 13.7 billion light-years, but don't forget that all the while the Universe is expanding, so the distant galaxy could now be a lot further than 13.7 billion light-years away. So, there appears to be a limit to the region of the Universe that you could have communicated with, even starting close to the Big Bang.

You might hope that you can get around that limit by starting much closer to the Big Bang, when everything was much closer together, but it turns out that the Universe was expanding much faster then, and that counteracts the advantage of starting earlier. Regardless of how soon you send your signal after the Big Bang, the distance the signal would travel by today has a maximum limit. (Astronomers call this limiting distance the *particle horizon*, though this book won't use that terminology besides this mention.)

This maximum limit you could have sent a signal is about 48 billion light-years, or about 15 billion parsecs. This is the distance from you today that the signal would have reached in the 13.7 billion years of the history of the Universe (it's more than 13.7 billion light-years because the Universe has been expanding in the meantime). This distance is also the furthest away that someone else could be for *you* to receive a signal from *them*. For this reason, this maximum distance is often referred to as **the size of the observable Universe**. Don't confuse this with the size of the entire Universe itself – one must carefully distinguish the *entire* Universe from the *observable* part of it, which could be a very small part. (In practice, one can't use light to look further than the surface of last scattering, which is a bit closer than the size of the observable Universe, but there are other signals that could penetrate the fog, such as using ripples in space and time called gravitational waves, or subatomic particles known as **neutrinos**. For that reason, strictly speaking one shouldn't use the cosmic microwave background to define the size of the observable Universe.)

- If the cosmic microwave background is now 48 billion light-years away, yet the Universe is only 13.7 billion years old, does that mean that it's been moving away faster than the speed of light?

  Having the space between two objects stretching is different from having one object in an unexpanding space moving away from another object. In an unexpanding space, motion is limited to less than the speed of light, but there aren't such restrictions for the stretching of space itself. Some science fiction writers use this theoretical possibility and imagine a 'warp drive' to get quickly from one place to another, without violating the speed of light, by stretching and squeezing space.

At any time in the history of the Universe, the observable Universe (or equivalently the region with which you could have been in contact) will have a particular size. For example, at the time that the Universe was last opaque (i.e. the time that the cosmic microwave background light was last scattered), this size was a lot smaller than today: about 1.5 million light-years, or 0.46 Mpc. That's closer than the Andromeda galaxy is to us today.

Einstein's theory of gravity, general relativity, together with the current knowledge of the contents of the Universe, can be used to calculate how fast the Universe is expanding, and how the expansion rate is changing. This in turn can be used to calculate how big an object 1.5 million light-years across would look. (You have to take into account the expansion of the Universe, because things appeared closer earlier on.) For example, at the time of last scattering of the cosmic microwave background, the size of the observable Universe on the sky would be about two degrees across, i.e. about four times the apparent diameter of the Moon. This means that if you were looking at a person in the microwave background, his or her observable Universe would look about two degrees across, as seen by you.

But this leads to a shocking conclusion. according to Einstein's theory of gravity, regions of the cosmic microwave background that are separated by more than two degrees *could not have been in any communication with each other*. So how is it that the cosmic microwave background ended up looking so extremely uniform? It should be extremely *non*-uniform. This is known as **the horizon problem**. This is a deep question that is surely saying something very important about the earliest moments of the Universe. One possible answer to this paradox will be explained in this chapter.

4.4 The curvature of space and the flatness problem

Another deep question is the flatness problem. In Einstein's theory of gravity, general relativity, space can be curved. (In fact, it's not just space that can be curved in this theory, it's also a more general thing called *spacetime*, but this book won't go into this distinction.) It's very hard – perhaps impossible – to imagine a three-dimensional space that's curved, so to imagine it one tends to use the same trick used before with a balloon: you would visualise the curvature of a two-dimensional surface, instead of a three-dimensional one. (People generally don't even try to imagine curved spacetime, as opposed to space.)

This is another example of how physics taxes the imagination. Our brains don't appear to be set up for imagining curved three-dimensional space plus time. This isn't the only difficulty with imagining things in physics. Even electricity and magnetism can be difficult, imagining the electric and magnetic components of electromagnetic waves, but electromagnetism is fundamentally connected to the theory of relativity, and the spacetime of relativity on its own is already demanding to imagine. If you go on to study quantum mechanics, you will probably find your mind recoils that the Universe is so strange, with waves of probabilities instead of solid objects. The Nobel-prizewinning physicist Richard Feynman once said: '… our imagination is stretched to the utmost, not, as in fiction, to imagine things which are not really there, but just to comprehend those things which *are* there'. Some people (physicists, naturally) have argued that physics makes the greatest imaginative and conceptual demands of any intellectual discipline.

Space can have almost any shape, but if space is homogeneous and isotropic, then Einstein's theory predicts that the large-scale shape of the Universe can take only one of three basic shapes These are quite hard to describe in three dimensions, but in two dimensions these forms are known as flat, spherical, or saddle-shaped (also called *hyperbolic*). If you zoom in close enough on any of them, they will all look like flat surfaces, but they look very different from a distance. Figure 4.2 shows these three types of surface.

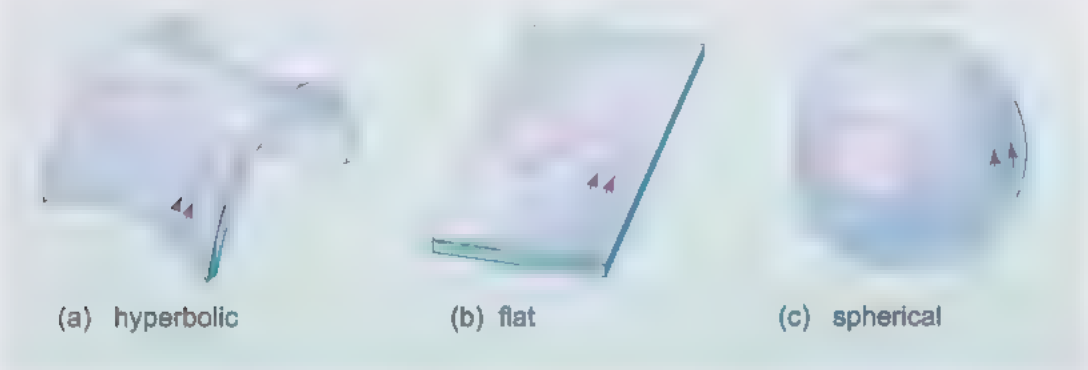


Figure 4.2 Two-dimensional representations of surfaces of different curvature. The left-hand surface (a) is saddle-shaped, or hyperbolic; the central surface (b) is flat, and the right-hand surface (c) is spherical. Lines that are initially parallel have different behaviours on these surfaces: in hyperbolic space, parallel straight lines diverge.

One way of imagining a hyperbolic space is shown in Figure 4.3. In two-dimensional flat space, it's possible to completely tile the surface using identical squares (e.g. a chessboard). Similarly, in three-dimensional flat space, it's possible to fill all of space with identical cubes. In a three-dimensional *hyperbolic* space, a dodecahedron (an object with 12 identical faces, each face being a regular pentagon) can be used to tile the whole of the space! Figure 4 3 shows a simulation of what such a tiling would look like from inside. Of course this is impossible in a flat space.

Another curious property of curved spaces is the behaviour of parallel straight lines. Now, when space is curved one has to be careful about what is meant by a straight line. Imagine a line of army ants on a two-dimensional surface. The surface can be as crinkly as you like, but if you zoom close enough in it should look flat. These army ants are small enough that they believe the surface to be flat. This army-ant-line will snake all over the crinkly surface, but the ants are so tiny they still believe they are fo lowing a straight line.

- Could these army-ant-lines ever cross themselves?

  Yes, depending on the shape of the surface. The ants are so tiny that they would still believe they're following a straight line, though.

These army-ant-lines are known as **geodesics** and are the curved space equivalents of straight lines. Now, the behaviour of these geodesics can be quite strange. In a flat space, straight lines that are parallel always stay

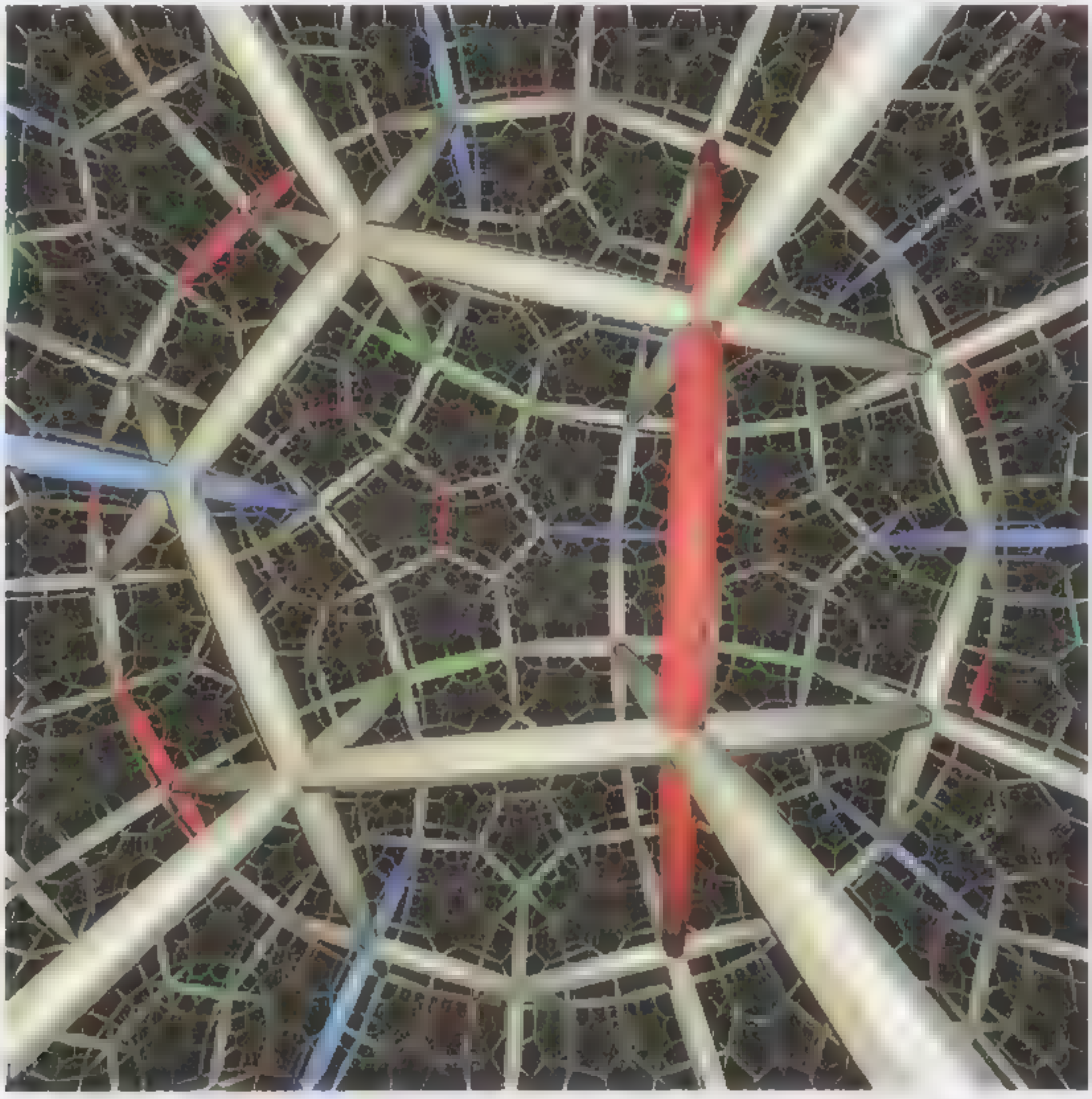


Figure 4.3 The view from inside a hyperbolic space, tiled with dodecahedron frames. This tiling is impossible in flat space.

parallel. This is shown in Figure 4.2. However, in curved spaces, geodesics behave differently. On a sphere, the geodesic lines are known as 'great circles', and the Equator is an example. Another example of a great circle on a sphere is a line of constant longitude, like the meridian line at longitude = 0 that passes through Greenwich, up towards the North Pole and down towards the South Pole. Two geodesics on a sphere can start off parallel, but they eventually meet. An example would be two nearby lines of constant longitude. At the Equator they are parallel, but they meet at the North and South Poles.

In hyperbolic space, parallel lines have the opposite behaviour: they diverge. These possibilities are all illustrated in Figure 4.2. How quickly the parallel lines diverge or converge tells you how curved the space is, and whether it's hyperbolic or spherical. There are other curiosities too: for example, the angles of a triangle add up to 180° in flat space, but add to less than 180° in a hyperbolic space, and to more than 180° in spherical space. The circumferences of circles are strange too: for a fixed radius, the circumference is bigger in hyperbolic space, and smaller in spherical space.

You may now have noticed that the balloon metaphor for the expanding Universe (Section 3.4) has a disadvantage: it's assuming the Universe has spherical spatial curvature. This is not necessarily the case. If the Universe is flat, or hyperbolic, then the expanding balloon picture is wrong. The replacement picture isn't quite so easy to visualise (an expanding saddle) and it also suggests space is infinite in size, because it can't have edges.

(At the risk of being confusing, there *is* a way for a flat or hyperbolic space to be not infinite in size. Have you ever played a video game where if you disappear off the right-hand side of the screen, you reappear on the left? Perhaps the Universe could be like that. If it is, then (it turns out) it's expanding too quickly for you to pull that trick off, unfortunately. In terms of how you'd imagine this expanding space, the expanding saddle metaphor would be wrong, because it doesn't have this curious wrap-around quality, but it's hard to come up with a better way of visualising it. Once again, it's very difficult to come up with accurate ways of visualising the physical Universe.)

But what *is* the spatial curvature of the Universe? Are we in a spherical Universe, as the balloon metaphor suggests, or a weird wrap-around saddle-shape, or something else? By measuring the contents of the Universe, and using Einstein's general relativity, it's possible to calculate the geometry of the Universe. It turns out that the Universe comes out as very flat. If there's any curvature (whether saddle-shaped or sphere-shaped), it's only apparent at sizes that are much bigger than can be measured in our observable Universe. Space could be curved and any size, so why does our observable Universe turn out to be so almost-perfectly flat? This is known as **the flatness problem**.

4.5 Inflation

This section will describe one possible answer that could solve the horizon problem and the flatness problem, as well as many other problems not discussed in this book. This theory is known as **inflation**. Now, many astronomers and cosmologists think it is the best theory that they have, but that doesn't mean they *believe* it is true (remember that ideally a scientist has no *beliefs* at all about any scientific theory). Many physicists are still actively keeping an open mind, or favouring an alternative theory as more likely in their view. While there isn't yet a consensus, it's still true that inflation has the broadest following.

The puzzle is that distant parts of the cosmic microwave background could not have been in contact, but somehow they end up almost exactly identical. How could this have happened? The approach in inflation is to change the calculation of the size of the observable Universe. This calculation used Einstein's theory of gravity, together with the current knowledge of the contents of the Universe. Inflation's approach is to suppose that some (as yet unidentified) content in the Universe triggered a much faster phase of expansion, known as *inflation*. The expansion history of the Universe would then look something like Figure 4.4.

What effect would this much faster expansion have? For one thing, it could explain why the space of our Universe looks flat. If you remember the crinkly surface with the army-ant-lines, the surface looked flat if you zoom in close

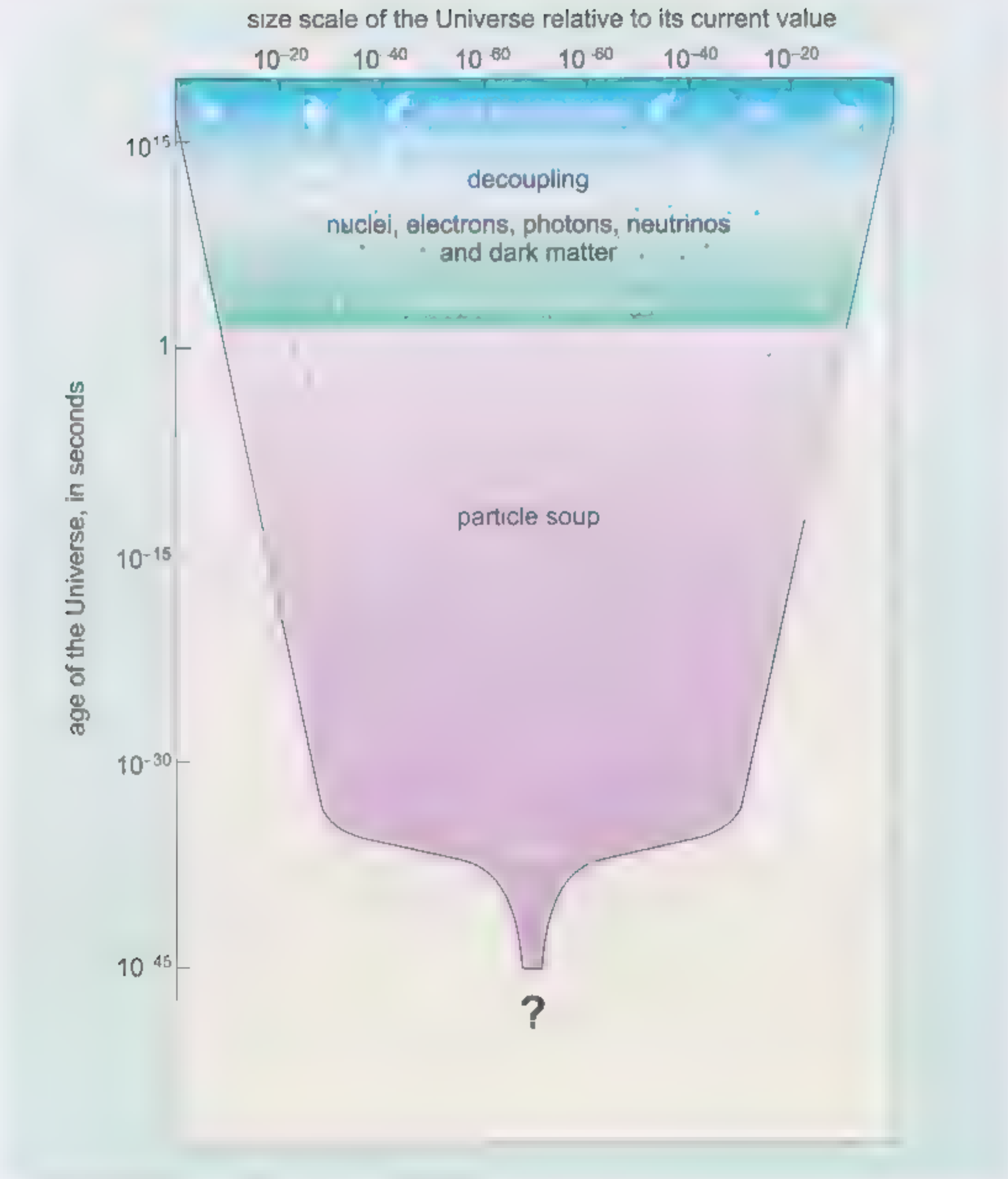

size scale of the Universe relative to its current value
10^-20
10^-40
10^-60
10^-60
10^-40
10^-20
age of the Universe, in seconds
10^15
1
10^-15
10^-30
10^-45
decoupling
nuclei, electrons, photons, neutrinos and dark matter
particle soup
?

Figure 4.4 The expansion history of the Universe.

enough. In inflation, the idea is that some tiny almost-flat portion of the Universe is expanded up to a giant size, which is why our observable Universe (or what we call our observable Universe) now looks so flat.

Another consequence is that scientists may have got the real size of the observable Universe wrong. In reality, at least according to the inflation theory, a tiny region of the Universe had plenty of time to send signals from one end to the other, and it's this tiny region that's been expanded up to at least the size of the Universe that we see today, and perhaps even much bigger. This explains why the cosmic microwave background seems so uniform – the patches of the sky that seem to have been completely out of contact, in reality (according to inflation) were very close together before inflation flung them apart.

Having said that, be warned that astronomers still use the phrase 'observable Universe' to mean the predicted size *without* inflation, despite a fairly wide support for inflation theory. This is perhaps because the extreme redshift that any pre-inflation signal would have would make it undetectable by any known technology, regardless of the type of the signal. Also, it's not known exactly how long the phase of inflation lasted for, so there is not yet a good way to calculate the size of the observable Universe taking full account of inflation

But what triggered this violent phase of expansion? Unfortunately, here the inflation theory becomes speculative. The speculation is possible because physicists don't know how the Universe works at very high energies – specifically, the energies of so-called **grand unified theories** that aim to describe three of the four fundamental forces of the Universe (electromagnetism, and nuclear forces known as *strong* and *weak*) with a single unified set of concepts. There is still uncertainty over how these grand unified theories (or **GUTs**) work. You may be surprised that such speculations are even possible. If so, it's time to reveal to you the dirty secret of physics. If you had signed up to study a foreign language, or engineering, or mathematics, you might expect that the very basics of the subject you are taught are at least logically self-consistent. In physics, we are sorry to tell you that there is no single self-consistent theory that describes all the experimental and observational data we have In particular, the theory of very large systems (Einstein's gravity theory known as general relativity) contradicts the theory of the very small (quantum mechanics). This contradiction isn't just at the level of specific predictions of experiments it's much worse. The contradictions are at fundamental conceptual levels. One speaks of *theories* in physics, such as the *theory* of relativity, and one hopes that as the theories describe more and more experiments correctly, they become closer and closer to the underlying laws of the Universe. Sometimes these theories are even given the name of *laws*, such as Newton's *laws* of motion, but if we took 'laws' to refer to the underlying laws of the Universe, these names would be merely vain boasts, because one couldn't possibly know for sure that we have got them right. And in fact physicists *know* that we are wrong in at least some fundamental ways, because their theories are contradictory. This is the dirty secret of physics, that despite huge successes from microchips to microwaves to spaceflight to nuclear power to the properties of materials to the colour of the sky, the subject is still not even logically self-consistent.

There are very few experimental constraints on how the Universe works at the energies of GUTs. Consequently there's still room for speculation at the GUT energies, without creating additional contradictions with known physical theories, and this is where the inflation theorists have made their stand.

The idea behind inflation is that for some (as yet not fully understood) reason, there was energy locked up everywhere in space. Then the Universe underwent a transition where this energy was released. This would be like the melting of ice, which releases latent heat. Melting ice is known as a **phase transition**, because the material changes from one phase (ice) into another phase (liquid water). The idea behind inflation is that some other phase transition in the early Universe released energy that drove the violent expansion of inflation. This rapid expansion cooled down the other contents of the Universe (remember from Section 4.2 that compressing a gas generally heats it up, while expanding a gas generally cools it down). Then at the end of the inflation period, the energy released from inflation stopped driving the expansion and instead went into heating the rest of the contents of the Universe back up again. This phase at the end of inflation is known as **reheating**. The exact timing of how long inflation lasted for, and what brought it to an end, both depend on the details of how the phase transition worked.

Inflation isn't proved yet – indeed you can't prove any theory in physics, only disprove it. Inflation passed its first test with the Cosmic Microwave Background maps of the NASA Cosmic Background Explorer (COBE) and the NASA Wilkinson Microwave Anisotropy Probe (WMAP), for which inflation made specific predictions about the fluctuations (Figure 4.1). The COBE results won Smoot and Mather the 2006 Nobel Prize in Physics. The next big test is a particular type of light called *polarised* light from the Cosmic Microwave Background. Regardless of how the phase transition worked, there are relationships predicted by inflation between the structures in polarised and unpolarised light. The European Space Agency Planck satellite is making these tests even as this sentence is being written.

Pushing even further back, what happened before the Big Bang? Unfortunately as you go further back beyond the inflation epoch, and the Universe was smaller and (consequently) hotter, you reach the sizes and energy ranges where quantum mechanics and general relativity clash in their fundamental concepts. Scientists simply don't know how the Universe works at sizes smaller than 10^{-35} metres, or on time intervals shorter than 10^{-43} seconds. This size scale is known as the **Planck length**, and the timescale is the **Planck time**. Both are named after the physicist Max Planck, who made pioneering discoveries in quantum mechanics. Winding back the history of the Universe, sooner or later you reach these limits.

Sometimes a glib answer is given to 'what happened before the Big Bang?': the Big Bang was when time started, so there's no 'before'. This is the situation that general relativity describes, but it doesn't explain the initial condition of the Universe of being flung apart. Moreover, this glib answer neglects to mention that scientists have no understanding at all of the very earliest moments, because the theory of general relativity contradicts the theory of quantum mechanics. Physicists have only informed speculations even at the epoch where grand unified theories are important. It remains possible that the theory of inflation is wrong, and it may yet be that the true answers to the horizon problem and the flatness problem lie at the limit of the Planck scales.

4.6 Optional section: what's the Universe expanding into?

What is the Universe expanding into? This is one of the most common questions asked at this level, so it has its own section. This section is marked as optional because other material won't relate to it, but it will give you an answer that is almost as far as human intuition can take you, which is unfortunately not very far.

The curvature of the *space* of our Universe was described in Section 4.4, but there have been several mentions of the fact that Einstein uses a more general concept called *spacetime*. This book won't go into the details of the meaning of this term, except to say that it's a more general way of regarding jointly both space and time. The Universe can have a flat *space*, but when you consider *spacetime* things become a bit more complicated. Incorporating time means one has to consider expansion or contraction of space too. It turns out that the expansion or contraction of the Universe can be regarded mathematically as being a curvature of *spacetime*. A flat space that's expanding can therefore nevertheless be regarded as a curved *spacetime*. This book tries to avoid curved spacetimes in order to prevent confusion. This will be the only section where they're referred to; apart from in this section, this book will only refer to the curvature and flatness of *space*.

The mathematics describing the curvature and expansion of the Universe can be made to work (this book won't go into any of this mathematics) but it's probably beyond the ability of human imagination to picture what the mathematics is describing. Therefore, let's go back to something that *can* be imagined: a two-dimensional curved surface. Even though this is a purely spatial surface (i.e. it's a space not a spacetime), the points to make about curvature will be true for spacetime too.

Firstly, let's go back to the army-ant-lines along a two-dimensional curved surface from Section 4.4. You may remember that these army-ant-lines are also known as *geodesics*. Recall that these are lines that look perfectly straight when you zoom in and look at them close up, but because of the curvature of the surface they end up wiggling all over the surface. If you wanted to make some measurements of the curvature of the surface, you might choose to take a step back and take some sort of three-dimensional measurements on your two-dimensional surface.

Einstein's theory uses a different approach to measuring the curvature of the surface. It turns out that all the curvature relevant to Einstein's theory can be calculated using *only* the geodesics. In the two-dimensional example, you only need to take measurements on the surface, i.e. how the army-ant-lines diverge or converge or cross each other; you don't *need* to refer to the three-dimensional space the surface is embedded in. In the case of our Universe, you only need to refer to measurements *within* our space and time – you don't *need* to embed our Universe in an impossible-to-imagine higher-dimensional space. You might still choose to, but the point is that it's not *required* by the mathematics that describes Einstein's theory. If the space of our Universe is curved, then you can measure that curvature within the space of our Universe, without needing to refer to any higher-dimensional space we're perhaps embedded in. Also, if our Universe is expanding (i.e. the *spacetime* of our Universe is curved), then one can still determine the expansion or spacetime curvature from measurements within our Universe. Again, it's not necessary to refer to a higher-dimensional space that our Universe might be embedded in.

So, one doesn't *need* to suppose that our Universe is embedded in a higher-dimensional space, i.e. the mathematics that describes the curvature and expansion of our Universe doesn't *require* that you embed our Universe in a higher-dimensional space. In other words, our Universe isn't (necessarily) expanding into anything at all – it's just expanding. If there is a higher-dimensional space 'out there' then we apparently can't communicate with it, so it's not part of our observable Universe.

This may not have left you convinced, but you may find another consideration helpful. If one is embedding our Universe in a higher-dimensional space, why should that higher-dimensional space be flat? At this point, any flatness or curvature has to be attributed to a physical cause.

It would be better if our brains were better able to imagine curved spaces. As it is, even experts in this area only have these imperfect analogies to work with; for the most part, the only other intuitive guidance even for the expert is the way the mathematical relationships work.

4.7 The fate of the Universe

A major result in the past few years has been to determine the ultimate foreseeable fate of the Universe, which will be described in this section. The content of the Universe is what determines the fate of the Universe. To see why, let's use the concept of **escape velocity**. How hard would you have to throw a ball to escape the Earth's gravity entirely? Neglecting air resistance, this comes out as 11.2 kilometres per second – far too fast to throw! If it's going more slowly than that, the ball would reach a maximum height then come back down. This is illustrated in Figure 4.5.

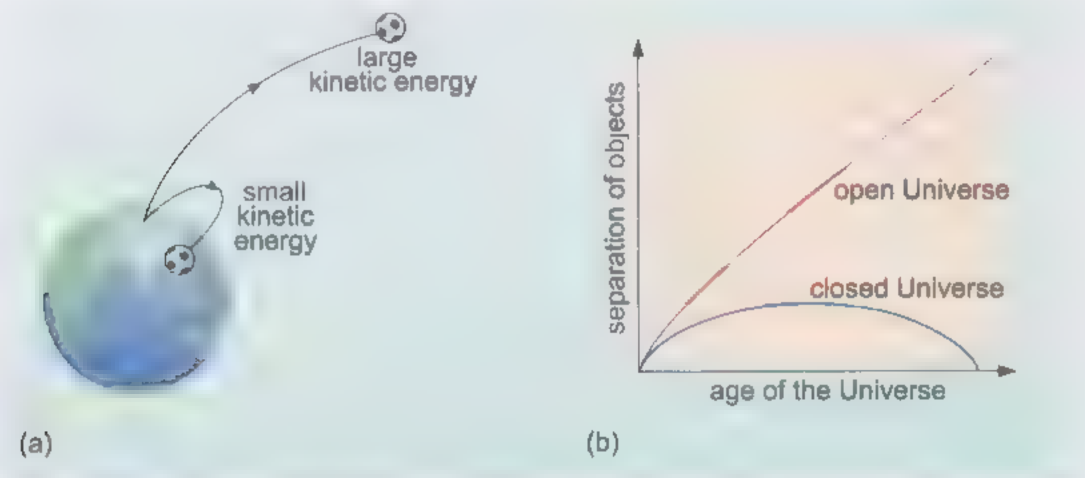


Figure 4.5 (a) Two possible paths for a ball escaping a planet's gravity (this is assuming no air resistance). If the ball has a high enough velocity (large kinetic energy), it can escape entirely. (b) The way that separations between objects change with time, depending on the fate of the Universe. In an open universe (as in being above the escape velocity), the separation increases continually with time. In a closed universe (as in being below escape velocity), a maximum distance is reached and then separations decrease.

Now, there's a similar thinking with the expansion of the Universe. Is there enough matter in the Universe for the gravitational attraction to stop the expansion, and bring it back into a big crunch? This is a bit like the escape velocity question, except this time we're keeping the velocity constant and asking if there's enough mass to bring the ball back down. This is also illustrated in Figure 4.5.

Chapter 3 had the story of how Einstein inserted a cosmological constant into his equations, which gave space an in-built tendency to expand This meant that he could describe the Universe as eternally held in a fine balance between the tendency to contract from gravity and the tendency of space to fling itself apart. Edwin Hubble took Einstein's original equations at face value and sought for the expansion or contraction of the Universe, and his discovery of the expanding Universe must surely rank as one of the greatest scientific discoveries ever made.

However, this is not where the story ends. The Universe sprung another surprise. By the 1990s, astronomical instrumentation had improved to the point that it became possible to make surveys of the sky for distant supernova explosions. A subset of supernovae known as type Ia (pronounced 'type one A') are particularly useful, because they have extremely characteristic luminosities. (These are 'standard candles' as described in Section 2.5.4.) This means that astronomers know exactly how much light they put out each second. This can be compared to how bright they *appear* on the sky, and that can be used to work out how far away they are. When this is compared to the measured value of the redshift, one can work out how much the Universe has expanded in the meantime (because it's the expansion that generated the redshift), so one can figure out how much the expansion of the Universe is slowing down. From that, it should be possible to figure out the fate of the Universe, because we'd be able to distinguish the 'below escape velocity' situation from 'beyond escape velocity'.

By the late 1990s the answer became clear from distant supernovae. The expansion of the Universe is not slowing down at all, but is in fact speeding up! This astonishing result won its discoverers (in practice, leaders of large teams) the 2011 Nobel Prize in Physics What could be causing this? The first and most obvious interpretation was to reprise Einstein's cosmological constant. This time though, it's not being used to hold the Universe in a steady-state 'just-so' balance between expanding and contracting. Instead, it's accelerating the expansion. If this theory is correct, the fate of the Universe is to expand faster and faster, with matter becoming more and more sparsely distributed in space.

There's an alternative way of looking at the cosmological constant, and to explain this we need to show you broadly how Einstein formulated his theory of gravity, general relativity. In Einstein's theory, all the matter and energy in space causes curvature of space (or more accurately, his more general concept of *spacetime*). The physicist John Wheeler is said to have expressed this relationship like this: 'matter tells space how to curve, and space tells matter how to move'. Einstein's fundamental equation is broadly of this form:

$$\left(\begin{array}{c}\text{something that}\\ \text{measures the}\\ \text{curvature of}\\ \text{space and time}\end{array}\right) = \left(\begin{array}{c}\text{something that}\\ \text{measures energy}\\ \text{and matter within}\\ \text{space and time}\end{array}\right)$$

Already, this is a very strange equation. What does it mean to set these two very different things equal? You could interpret this as saying that energy and matter are telling space how to curve, as John Wheeler described. However you could equally read the causal link the other way, and regard curvature as somehow causing the mass and energy. The origin of mass in particle physics is a very deep problem and one that the Large Hadron Collider aims to illuminate, but understanding the link with spacetime curvature may well be beyond its reach. Einstein himself is said to have remarked that the left-hand side of this equation is carved in marble (because this is the beautiful geometry of his theory), while the right-hand side is made of straw (because there is no compelling answer as to why matter and energy should have this effect).

Now let's add the cosmological constant, as Einstein did. The symbol used for the cosmological constant is the Greek capital letter lambda, or Λ. Einstein's modified equation is of this form:

$$\left(\begin{array}{c}\text{something that}\\ \text{measures the}\\ \text{curvature of}\\ \text{space and time}\end{array}\right) + \Lambda = \left(\begin{array}{c}\text{something that}\\ \text{measures energy}\\ \text{and matter within}\\ \text{space and time}\end{array}\right)$$

So, by putting the Λ on the left-hand side, this is regarding the cosmological constant Λ as a property of space and time. However, the cosmological constant Λ is just a number, so one could subtract Λ from both sides and the equation would still be true. The + Λ − Λ on the left-hand size would become zero (because anything minus itself is zero), so the equation would then look like this:

$$\left(\begin{array}{c}\text{something that}\\ \text{measures the}\\ \text{curvature of}\\ \text{space and time}\end{array}\right) = \left(\begin{array}{c}\text{something that}\\ \text{measures energy}\\ \text{and matter within}\\ \text{space and time}\end{array}\right) - \Lambda$$

This time, Λ is regarded as something within space and time, rather than a property of space and time itself. What sort of substance would Λ represent? A very strange one, it turns out. It would have a negative pressure, and no everyday substance has that property. One can also regard Λ as representing an energy density, i.e. an amount of energy for every cubic metre of the Universe. It turns out that this energy density is more than all the matter in the Universe put together in fact, over 70% of the energy density of the Universe is from this Λ term. Yet scientists have no idea why there should be a Λ term, nor why it should take the numerical value that it does. Some attempts have been made to predict the numerical value of Λ from quantum mechanics, but they get the wrong answer by an astonishing factor of 10^{120}, i.e. a factor of

1 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000
000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000.

This is perhaps the worst disagreement of experiment with best-available theory that there has ever been in the history of all science, yet physicists still have no better prediction for the numerical value of Λ.

Some have wondered if a different approach is needed. Empirically, people have considered a generalisation of Einstein's cosmological constant. In Einstein's theory, Λ has a constant value that is the same everywhere and at every time in the Universe. But one might choose to go beyond Einstein, and start to suppose that Λ could behave differently in different places, or at different times. This is what is known as **dark energy**. It's still being seen as a substance filling all space and time, but now it can vary in strength from time to time and from place to place. At the time this is being written, there is currently no evidence that dark energy is anything different from the cosmological constant that Einstein originally envisaged. However, many sensitive experiments are underway or are planned to see if different lines of sight give different values for the acceleration of the Universe's expansion, or whether the changes in the acceleration exactly match the predictions from Einstein's theory. If these experiments find evidence for dark energy that differs from the cosmological constant, this would be an immensely important discovery, perhaps revealing something fundamental about how space and time themselves are built. In 2010, NASA had a once-in-a-decade review of science priorities, and ranked an experiment to measure the effects of dark energy as its top priority.

Now, this section has mentioned dark energy, and dark *matter* was mentioned in Chapter 3. It's very easy to get these confused, so it's important for you to understand that they are very different things.

Dark matter is not dark energy!

Dark *matter* collects in clumps known as haloes, and has a visible effect on how galaxies rotate, and how they move relative to each other (especially in clusters of galaxies). If you added more dark matter to the Universe, the additional gravitational attraction would *slow down* the expansion of the Universe.

Dark *energy* does not clump (as far as we've been able to tell so far) and its principal effect is to *speed up* the expansion of the Universe.

However, dark matter and dark energy do have one thing in common: astronomers don't know much about either of them. Scientists have no definite idea what the fundamental particles of dark matter are, so far at least. They also have no idea whatsoever why the Universe should have dark energy in it, or whether it's just the cosmological constant that Einstein originally envisaged. Physicists can't even get remotely close to a prediction of the value of Λ. Figure 4.6 shows a pie chart of the different contributions made to the energy density of the Universe. It turns out that the total from all these contributions is closely related to the large-scale curvature of the Universe (i.e. spherical, flat or saddle-shaped). Most of the energy density in the Universe is dark energy (or cosmological constant). The next largest contributor is dark matter. It's a surprising - and perhaps shocking state of affairs that nearly all of the contents of the Universe are so poorly understood.

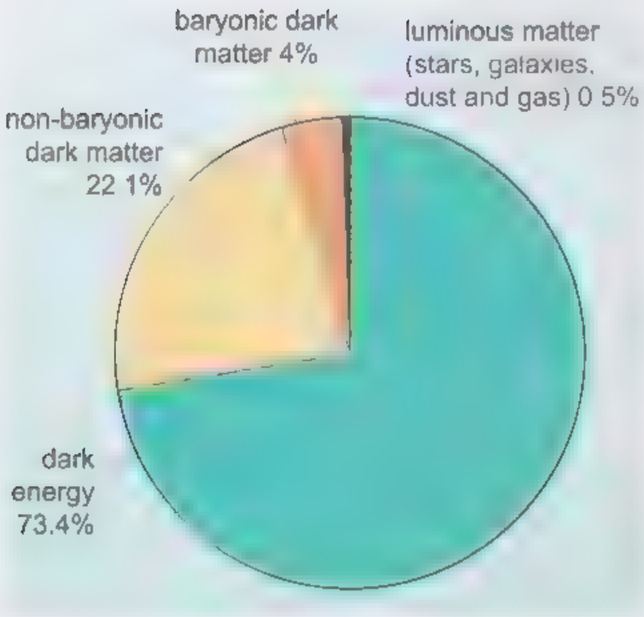


Figure 4.6 The relative contributions to the energy density of the Universe. Most of the energy is in the form of dark energy. The next largest component is dark matter, made of collisionless dark matter particles (meaning that they pass through each other and have no turbulence or air resistance). There is also a contribution of non-luminous matter that's made of the same particles as everyday objects, known as baryonic dark matter Finally, there is the luminous matter All the stars and all the gas in all the galaxies are a very small part of the total!

The final words in this section will be about the ultimate fate of the Universe. As the Universe expands, the matter becomes more and more sparsely distributed. The cosmological constant (or dark energy) then becomes increasingly important in driving the expansion, until it becomes the only important factor. How far ahead is it possible to look? It depends on what you choose to assume about dark energy. If it's the cosmological constant that Einstein envisaged, then the Universe would eventually settle down into a constant expansion. Perhaps a trillion years after the Big Bang, matter would eventually become too sparse for star formation to happen, so the Universe would eventually once again become dark. One can then look even further ahead to the decay of all the particles in matter (even those that currently appear stable in laboratories), but the difficulty is that astronomers don't have good observational constraints on dark energy or laboratory measurements of dark matter particle properties, so it's hard to foresee. This is the difficulty with very long-range forecasting of the Universe: eventually you reach the point that is beyond the knowledge of the physical Universe given by your laboratory constraints. The next section will show you how far you can go at the other end of the history of the Universe, the first 15 minutes of the Universe, without straying from the experimental constraints of laboratory physics.

4.8 What are all visible things made of?

There is a children's saying that girls are made of sugar and spice and all things nice, while boys are made of frogs and snails and puppy dogs' tails. Whatever the merits or problems of that metaphor, it's obviously not literally true Boys, girls, sugar, snails and all visible matter are made up of atoms, as you saw in Chapter 1. There's even atomic matter that's not obviously visible, as was shown in Figure 4 6, though most 'dark' matter is not atomic.

Perhaps all knowledge is a double-edged sword. Advances in nuclear physics have led to magnetic resonance imaging (MRI) in medicine, nuclear power and nuclear weapons. In 1948, three years after the atomic bombs were dropped on Hiroshima and Nagasaki, it was realised that the same fundamental theories of nuclear physics can be applied to the understanding of the early Universe. In order to explain the insights this revealed on the first fifteen minutes of the history of the Universe, you need to know a little about the structure of atoms.

In Chapter 1 we showed you that everything around you is composed of atoms. There are known to be around ninety different types of atoms that occur naturally in the world, and in Chapter 1 we introduced the term *element* to describe these different types of atom. The Universe as a whole is mostly composed of the elements hydrogen and helium, and this section will show you why. Here on Earth, there are also significant amounts of other elements, in particular carbon, nitrogen, oxygen, sodium, magnesium, aluminium, silicon, sulfur, calcium and iron. Some substances are pure elements. Diamond crystals, for example, contain only carbon atoms arranged in a regular rigid structure. Other substances take the form of chemical compounds. For example, common salt (sodium chloride) is another regular crystal structure containing both sodium and chlorine atoms. In gases or liquids, atoms may combine together to form *molecules* (another term from Chapter 1). You probably know, for instance, that the chemical symbol for carbon dioxide is CO_2. This indicates that a carbon dioxide molecule is composed of one carbon atom and two oxygen atoms. Similarly, ammonia has the chemical symbol NH_3 – the molecule contains one atom of nitrogen and three of hydrogen.

- Of what atoms is a water molecule (H_2O) composed?

One oxygen atom and two hydrogen atoms.

Just to remind you, we'll recap what atoms are made of. Each atom contains a central *nucleus*, which carries a positive electric charge as well as most of the atom's mass. The nucleus is surrounded by one or more negatively charged particles known as *electrons* (symbol: e^-) each of which has a much lower mass than the nucleus. The nucleus of an atom is what determines the type of element. The very simplest atoms of all, those of the element hydrogen, have a nucleus consisting of just a single particle, known as a *proton* (symbol: p). The positive charge on a proton is equal and opposite to the charge on an electron (see Table 4.1). The next simplest atom, helium, has two protons in its nucleus; lithium has three protons; beryllium has four; boron has five; carbon has six; and so on. The number of protons in the nucleus of an atom is known as its *atomic number*.

The other constituents of atomic nuclei are particles known as neutrons (symbol: n). Neutrons have a similar mass to that of protons, but have zero electric charge. They therefore contribute to the mass of an atom, but not to its electric charge. (Protons and neutrons are themselves made of smaller particles known as quarks, but they will not be discussed here.) Normal hydrogen atoms do not have neutrons in their nuclei, although there is a form of hydrogen known as **deuterium** that does. The nucleus of a deuterium atom consists of a proton and a neutron. It is still the element hydrogen (because it contains only one proton) but it is a 'heavy' form of hydrogen, thanks to the extra neutron. Deuterium is said to be an **isotope** of hydrogen. Similarly, normal helium atoms contain two neutrons in their nucleus, along with the two protons, but a 'light' isotope of helium contains only one neutron instead. The total number of protons *and* neutrons in the nucleus of an atom is known as the **mass number** of the atom.

Table 4.1 The constituents of atoms: subatomic particles.

	Electric charge	Notes
electron e	−1 unit	In a neutral atom, number of electrons = number of protons
nucleons:		*Mass number* = number of nucleons
proton p	+1 unit	*Atomic number* = total number of protons
neutron n	0	*Isotopes* of the same elements have different numbers of neutrons

■ What are the mass numbers of (a) normal hydrogen, (b) 'heavy' hydrogen (i.e. deuterium), (c) normal helium and (d) 'light' helium?

(a) The nucleus of normal hydrogen contains one proton, so the mass number is 1.

(b) The nucleus of deuterium contains one proton and one neutron so the mass number is 2.

(c) The nucleus of normal helium contains two protons and two neutrons, so the mass number is 4.

(d) The nucleus of 'light' helium contains two protons and one neutron, so the mass number is 3.

Isotopes of each atomic element may be represented by a symbol. Letters are used to indicate the name of the element itself, and two numbers are used to indicate the atomic number (lower) and mass number (upper). Hence a normal hydrogen atom is represented as $^{1}_{1}\mathrm{H}$, and an atom of the heavier isotope, deuterium, by $^{2}_{1}\mathrm{H}$. Isotopes of some other light atoms are indicated in Figure 4.7.

Normal atoms are electrically neutral, i.e. the positive electric charge of the nucleus is exactly balanced by the negative electric charge of the electrons surrounding it. Because each electron carries an electric charge of −1 unit (see Table 4.1) and each proton carries an electric charge of +1 unit, the number of electrons in a neutral atom is *exactly* the same as the number of protons in its nucleus.

Atomic nuclei are very small, typically around 10^{-14} metres. The size of an atom itself, however, is determined by the region occupied by the electrons that surround the nucleus. This is typically around 10^{-10} metres.

■ How much larger than a nucleus is an atom? Think of some everyday examples.

An atom is about $(10^{-10}/10^{-14}) = 10^{-10} \times 10^{14} = 10^4$ times larger than the nucleus (because dividing by something to a negative power is equivalent to multiplying by that same thing to its positive power (and vice versa), and $-10 + 14 = 4$). The continent of Australia is roughly 4000 km across. An atom this size would have a nucleus only 400 metres across, about the size of a city block. The M25 ring road around London has a diameter of roughly 40 km. If an atom were this size, the nucleus would be only 4 m across, or the size of a taxi in central London.

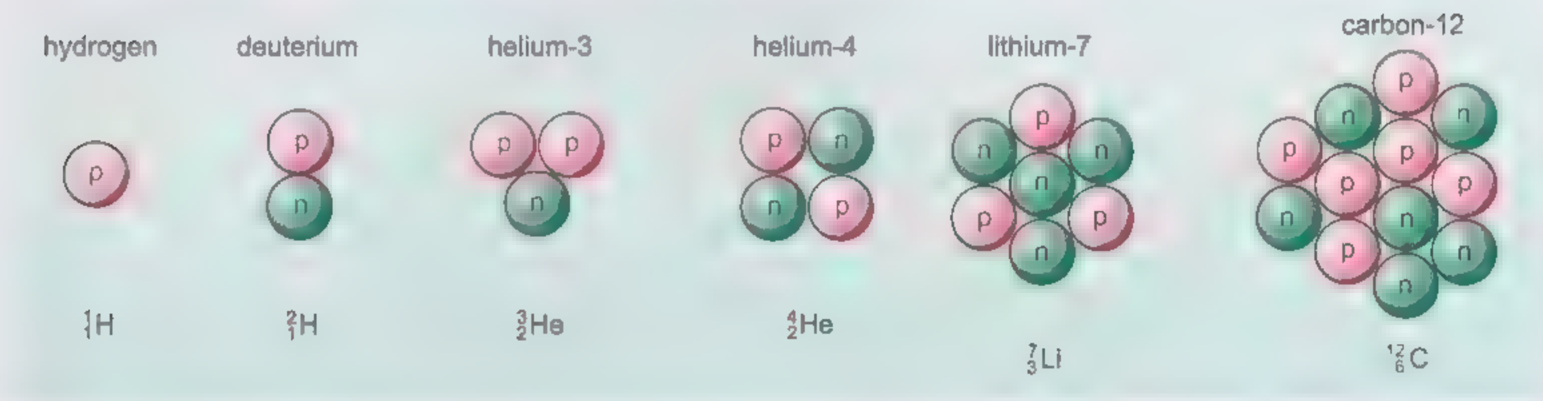


Figure 4.7 Schematic diagrams of the nuclei of some isotopes. Protons are coloured red and labelled with p, and neutrons green and labelled with n.

For a number of years at the beginning of the twentieth century it was believed that the electrons in an atom travelled around the nucleus following orbits, in much the same way as planets orbit the Sun in our Solar System, except the electrons aren't confined to a plane (Figure 4.8).

However, since the 1920s it has been known that this cannot be correct – the situation is much stranger. The 'planetary orbit' description of atoms in Figure 4.8 is good enough for all purposes in this book, but just so you aren't left with a misleading impression, here is a more accurate description.

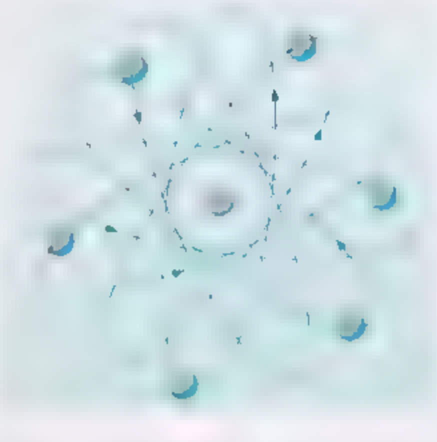


Figure 4.8 Electrons in orbit around an atomic nucleus. Although conceptually appealing, such a picture has been known to be *false* since the 1920s. A better representation is shown in Figure 4.9.

It is now known that electrons in atoms exist in a 'fuzzy cloud' around the nucleus, in which both their speeds and positions are indeterminate. That is, it is impossible to predict, in advance of a measurement, exactly where an electron will be found or what speed it will have. In some sense, each electron is actually in many places at once, and all that can be said is that there is a certain **probability** of each electron being in a particular location with a particular speed. This picture is predicted by **quantum physics** and has been verified experimentally. An **electron probability cloud** for the single electron in a hydrogen atom is shown in Figure 4.9 (overleaf). The cloud is dense in those regions where there is a high probability of finding the electron, and sparse in those regions where the electron is unlikely to be found. One way of thinking about such a cloud is as follows. Imagine that you could perform an experiment to measure the position of the electron within a hydrogen atom. If you did this ten thousand times, and each time drew a dot at the location where you found the electron to be, the result would be something like Figure 4.9. Of course, this is only a two-dimensional drawing of the real three-dimensional situation, but it gives the general idea.

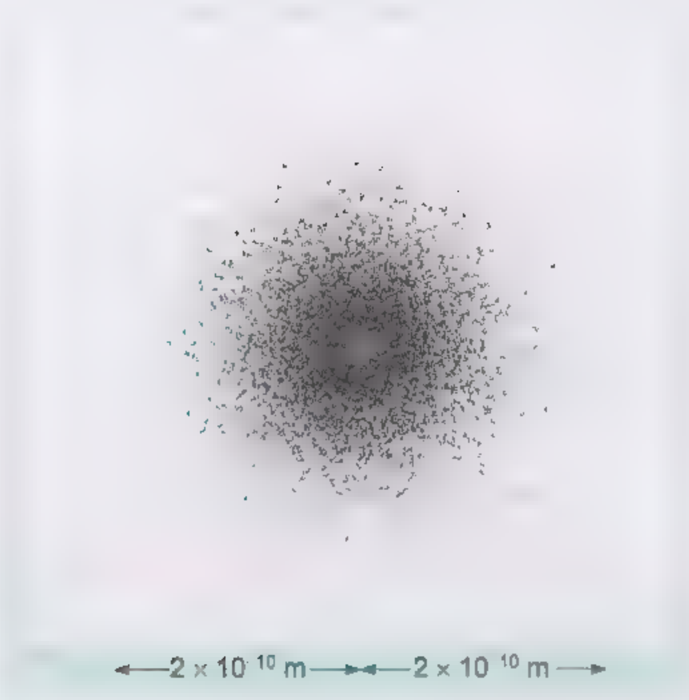
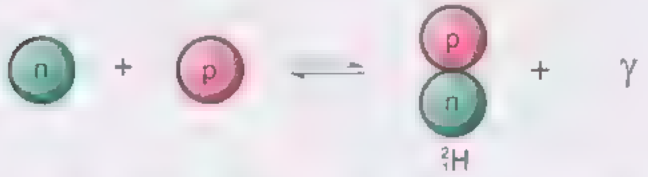


Figure 4.9 The electron probability cloud for the single electron in a hydrogen atom.

4.9 The first fifteen minutes of the Universe

Having given you this brief run-down of the structure of atoms, let's now take you back to the early Universe to show you the inferences that can be made, at times of between 100 seconds and 1000 seconds after the Big Bang. The end result will be an observation that is now regarded as one of the three or four greatest confirmations of the Big Bang theory.

About 100 seconds after the Big Bang, the Universe was much hotter and denser than it is now, with a temperature of about 10^9 K. The content of the Universe was largely dark matter particles, plus protons, neutrons, electrons, photons, and another type of particle you met in Section 4.3 known as neutrinos. As the Universe expanded the temperature decreased, and the protons and neutrons were able for the first time to combine and form light nuclei. This marked the beginning of the period referred to as the era of primordial **nucleosynthesis** (which literally means 'making nuclei'). The first such important reaction was that of a single proton and neutron combining to produce a deuterium nucleus, with the excess energy carried away by a high-energy photon known as a **gamma-ray**. This reaction is shown in Equation 4.1. The double arrows show the reaction can go either way.



(4.1)

At higher temperatures (greater than 10^9 K) there were a lot of high-energy photons so this reaction is favoured to go from right to left. As a result, deuterium nuclei were rapidly broken down. However, as the temperature fell below 10^9 K when the Universe was about 100 seconds old, deuterium production was favoured. Virtually all of the remaining free neutrons in the Universe were rapidly bound up in deuterium nuclei, and from then on other light nuclei formed. In one of the reactions that occurred the net result was the conversion of two deuterium nuclei into a single nucleus of helium-4.

Other more massive nuclei were made: almost entirely beryllium (which is unstable) and lithium. Neutrino particles were created in these reactions but are quite unreactive. A prediction is that there should be many of these neutrino particles left over from these primordial reactions. Neutrinos are however very difficult to detect, precisely because of this lack of reactivity. The primordial neutrinos have not yet been detected directly, but there is also a constant stream of neutrinos passing through you and through the Earth from the Sun; these neutrinos come from the nuclear reactions in the core of the Sun. These solar neutrinos are much more abundant in our neighbourhood and have been detected directly. There will be more about nuclear reactions in stars in the next chapter.

Nuclei with a mass number greater than seven did not survive in the early Universe. This is because there are no stable nuclei with a mass number of eight – notice from above that the beryllium nuclei decay spontaneously, leading ultimately to more helium-4. The reactions that would by-pass this bottleneck take much longer than the few minutes that were available for nucleosynthesis at this time. (Remember, this is a timespan of around 15 minutes when the Universe had an age of between 100 and 1000 seconds.) Before more advanced reactions could occur, the Universe cooled down too much to provide the energy necessary to initiate them.

Protons are very stable particles, but free-floating neutrons decay spontaneously into other particles. The ratio of protons to neutrons had, by this time, reached about seven protons for every one neutron. Because the neutrons were then bound up in nuclei, they no longer decayed, and the ratio remained essentially fixed from here on. The vast majority of the neutrons ended up in nuclei of helium-4. Only very tiny fractions were left in deuterium, helium-3 and lithium-7 nuclei, because the reactions to produce them were far more likely to continue and produce helium-4 than they were to halt at these intermediate products.

By the time the Universe had cooled to a temperature of about 3×10^8 K after 1000 seconds, the particles had insufficient energy to undergo any more reactions. The Universe was only about a quarter of an hour old, but the era of primordial nucleosynthesis was at an end, and the proportion of the various light elements was fixed. The rates of reaction to form helium and the other light elements have been calculated, and the abundances predicted may be compared with the abundances of these nuclei that are observed in the Universe today. The proportion by mass of helium-4 that is predicted to have come out of the Big Bang is about 24.8% and the observed value is between 23.2% and 25.8%. There is clearly close agreement between theory and observation. The proportions by mass of deuterium and helium-3 to emerge from primordial nucleosynthesis were only about 1 in 40 000 and 1 in 100 000 respectively, and the proportion by mass of lithium-7 was about 1 in 2 billion.

In 1948, the PhD student Ralph Alpher and his supervisor George Gamow first proposed calculating the reaction rates of primordial nucleosynthesis. (As a joke, they managed to get the agreement of their colleague and friend Hans Bethe to co-author the paper, so the author list read *Alpher, Bethe, Gamow* mimicking the first three letters of the Greek alphabet.) The initial reaction to this paper was hostile. What made the authors believe they could apply the physical theories of nuclear physics so soon after the Big Bang? However, their work was applying well-tested laboratory physics, and in the end the Universe provided the corroboration: the observed abundances of hydrogen and helium in the present-day Universe are in strikingly good agreement with the predictions of primordial nucleosynthesis. The agreement is so good that this is regarded as one of the main lines of evidence supporting the Big Bang theory itself, along with the expansion of the Universe and the cosmic microwave background.

At an age of 1000 seconds, the Universe reached a state where its matter constituents were essentially as they are today. There are about 10 billion photons for every baryon (proton and neutron), and about seven protons and electrons for every one neutron. Neutrinos continue to travel through the Universe unhindered by virtually anything they encounter.

These weren't the last nuclear reactions that happened in the Universe; nuclear reactions are also responsible for the energy output of our Sun and other stars. The observed abundances of heavier elements are harder to compare to the primordial predictions, because of the element production that happens in stars. However, this heavy element production in stars is responsible for everything you see around you right now: almost all the elements in rocks or the air or in living things, in fact pretty much all the visible matter in the Universe bar hydrogen and helium, has been generated in stars. The next chapter will cover the nuclear reactions in stars that have been so crucial to the formation of heavier elements in the Universe, without which there could have been no planets and no life.

End-of-chapter questions

Question 4.1 Suppose that as soon as you were born, you made a phone call on a satellite phone Ignoring the fact that you'd have been amazingly precocious, calculate how far in metres that radio signal would have travelled in your lifetime up to now.

Question 4.2 Assume that the Universe contains one neutron for every seven protons, that all the neutrons are today bound up in nuclei of helium-4 and that the mass of a proton is about the same as the mass of a neutron.

(a) What are the relative numbers of hydrogen and helium nuclei in the Universe?

(b) What are the relative percentages, by mass, of hydrogen and helium in the Universe?

Question 4.3 Write a few sentences in your *own* words describing the various contributions to the matter and energy densities of the Universe and comparing them qualitatively.

Question 4.4 Write a few sentences in your *own* words describing what the cosmic microwave background is, and why it is now 48 billion light-years away, despite the fact that the Universe is only 13.7 billion years old.

Now go to the module website and do the remaining activities associated with Chapter 4.

Activity 4.1 - 30 mins

Activity 4 2 - 30 mins

Activity 4.3 - 20 mins

Activity 4.4 - 20 mins

Chapter 5 The structure and life of stars

At the end of the previous chapter we stated that stars played a crucial role in how the chemical composition of the Universe changed over time, due to the production of heavier elements that are found in the Earth and indeed in all of the life on it. But how does this process work – what physical processes lead to this nucleosynthesis and where do they predominantly occur? Answering these questions leads to an investigation of the life cycle of stars. The next two chapters focus on their lives, births and deaths. The best place to start this journey is with the nearest star to us – the Sun. The reason for this is simple – at a distance of 'only' 150 million kilometres it is roughly 270 000 times closer than our next nearest stellar neighbour. Consequently, while all but a handful of other stars are so distant they appear as point-like pinpricks of light to even our largest telescopes, astronomers can study the surface of the Sun and hence infer the behaviour of its interior – in exceptional detail.

5.1 Our nearest star: the Sun

Chapter 1 discussed the basic appearance of the surface of the Sun (photosphere) and its extended atmosphere (chromosphere and corona). The Sun is a highly dynamic and variable object To account for its brightness and activity, the Sun must contain a power source. However, the nature of that power source was a great puzzle until the early twentieth century. Fossil records and ideas about evolution were beginning to provide firm evidence that the Earth must be at least hundreds of millions of years old, rather than thousands of years as was previously thought, and the Sun must be at least as old as the Earth. The only fuels known at the time were coal, wood, oil, gas, and so on. It was fairly easy to calculate that, even if the Sun were made entirely of one of these fuels, and could get the necessary oxygen from its surroundings, it could burn for only a few thousand years at most while producing its current output of heat and light – not nearly long enough to sustain life on Earth over millions of years.

The problem of the Sun's fuel baffled many of the world's best scientists until **nuclear reactions** were discovered in the early twentieth century. Such reactions provided a totally new type of energy source. Rather than burning like coal or gas, nuclear reactions need no oxygen and produce vastly more heat and light for a given amount of fuel. Nuclear reactions give 'atomic' weapons their great destructive power but are harnessed more productively in electricity generation. The type of reactions that power the Sun – so-called *fusion reactions* involving hydrogen – are similar to those that occur in a hydrogen bomb but in the Sun they proceed steadily rather than as an explosion.

The British astronomer Arthur Eddington calculated that, if the Sun were made mainly of hydrogen undergoing nuclear reactions, it could last for millions of years while producing a more-or-less steady heat and light output. Furthermore, its outward appearance would closely resemble that of the actual Sun. It's now known that hydrogen nuclear reactions will sustain the Sun for about ten thousand million years. The Sun is currently about half-way through the hydrogen-fuelled phase of its life.

Everything that is known about nuclear reactions is based on experiments performed in laboratories on Earth. Eddington's great triumph was being able to take that knowledge and work out what would happen if nuclear reactions happened on a far larger scale than was possible on Earth, and to relate his deductions to what was known about the Sun. This example illustrates an important feature of astronomy. everything that is known about the Universe beyond our Earth and Moon (apart from a few planets that space probes have visited) must be deduced by observing from a very great distance. Astronomers have two main strands to their quest to understand such distant objects. One strand involves the observations themselves: studying the appearance of distant objects and analysing the radiation received from them The other strand involves finding out how objects behave on Earth and using that knowledge to interpret and account for the observations.

Scientists have used what they know about nuclear reactions and about how very hot materials behave, together with detailed observations of the Sun, to piece together a model (that is, a mental picture) of what the Sun must be like deep inside.

This is shown in Figure 5.1. The Sun does indeed consist largely of hydrogen and it is gaseous throughout. The nuclear reactions occur only in the Sun's **core** - that is, deep in its centre because the hydrogen fuel needs to be at a temperature of over 10 million K before nuclear reactions can begin. Energy is carried away from the core by radiation that is repeatedly absorbed and re-emitted as it travels through the **radiative zone**. Initially, the energy released is transported outwards as high energy photons known as gamma-rays (you may remember them from Chapter 4), but by the time it reaches the surface, the process of constant absorption and re-radiation has led to a decrease in the average energy of each photon and they are emitted from the Sun at optical wavelengths. The total energy released in the core of the Sun is still the same as is radiated at its surface.

Before the photons leave the surface of the Sun, the energy is transferred by a different process, known as **convection**. Like radiation, convection is another way that energy can be transferred from a hot to a cool region. In radiation, no matter moves and energy is transferred by photons. In convection, the material itself moves, carrying the heat energy with it. This can happen when a layer of hot gas or liquid lies beneath a layer of cooler gas or liquid.

■ Can convection happen in a solid?

□ Convection involves the motion of material so it cannot occur in a solid.

The hot material is less dense, so it's more buoyant than the cooler material and rises through it. Hot air in a room rises in the same way. The cooler material sinks down below the hotter layer. If there is a constant source of energy, this cooler material then heats up. The cycle then repeats itself and a **convection cell** is set up with material constantly being heated and rising before cooling and sinking. This process is surprisingly common in nature and occurs in places as diverse as the interior and atmosphere of the Earth and in a Lava Lamp!

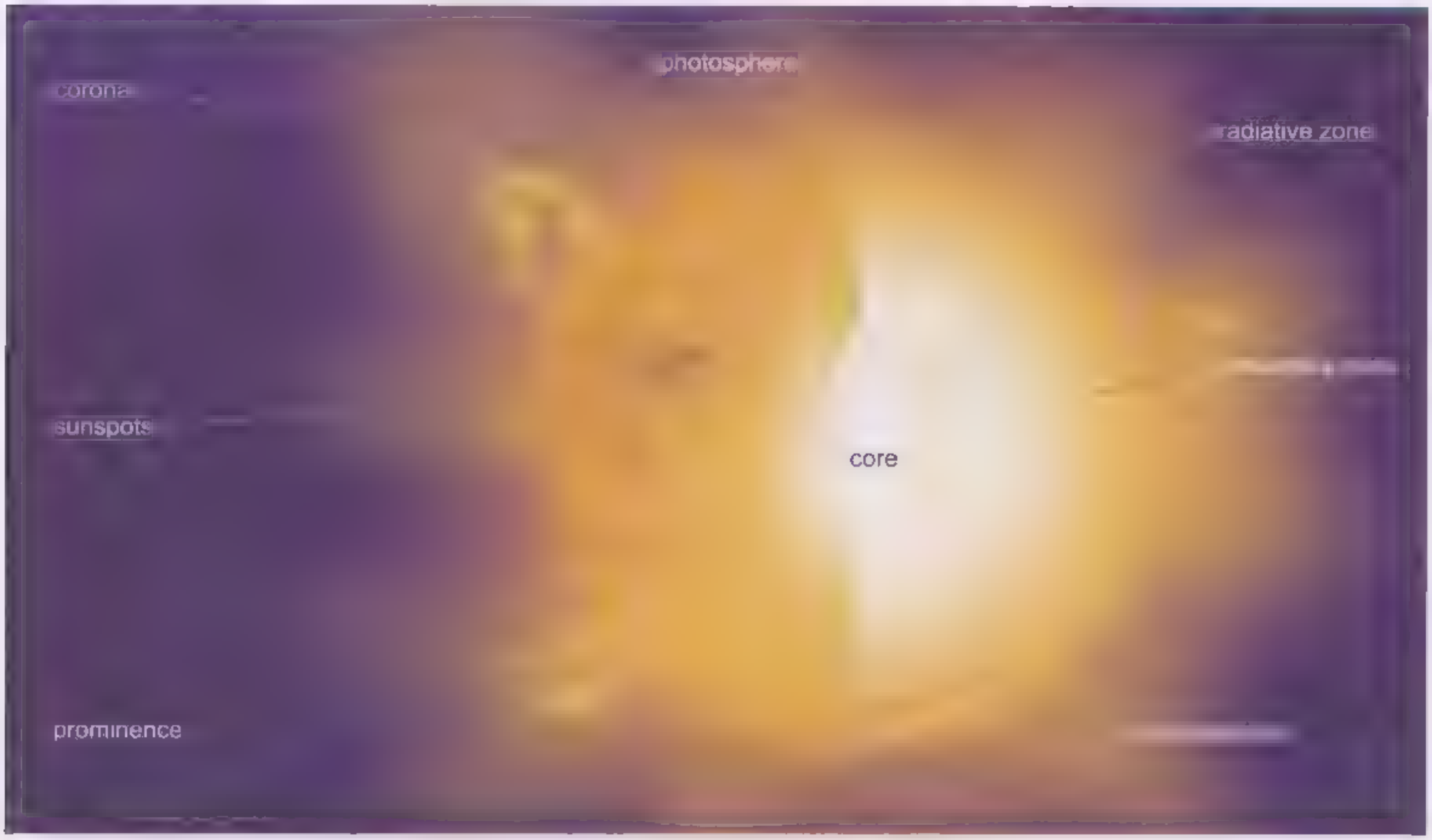


Figure 5.1 The solar interior Temperature and density increase rapidly with depth inside the Sun, but only in the central core are the conditions right for nuclear reactions to occur. Beyond the core there are zones where energy is transported to the surface by processes involving radiation and convection.

Activity 5.1 Convection in action

The estimated time for this activity is 10 minutes

The easiest way to see convection in action is to fill a saucepan quarter-full with water and then place it on a cooker hob and gradually heat it. (You should NOT attempt this activity using a microwave or open kettle.)

As the water heats up you'll see it start to bubble and churn with the motion of the water transferring the heat from the bottom of the pan to the surface of the water. Indeed the churning pattern of granules on the surface of the Sun is the visible manifestation of convection, being the top of convection cells. Do you see patterns of convection cells in your pan? Do they correspond to the heat from underneath the pan?

This process occurs in the **convective zone** inside the Sun. As the rising columns of hot material approach the top of the convective zone, the material above becomes thinner, increasing the chance of any emitted light escaping into space. This 500 kilometre thick region constitutes the photosphere, and the rising and falling columns of solar material account for the seething pattern of granules mentioned in Chapter 1 (Figure 1.8b).

But why is the core this hot? Why does it need to be so hot for nuclear reactions to happen? And how do these deep interior processes affect how the Sun appears from the outside? The rest of this chapter will answer these questions.

At this point, you might well be wondering exactly how anyone can know so much about the Sun's interior when astronomers can only see its outer layers. Astronomers have used two techniques to address this problem one is based on understanding the physical processes that power the Sun (you will meet these later on in this chapter) and the second **helioseismology** – is borrowed from how geologists study the Earth. Since the 1960s astronomers have known that the surface of the Sun is constantly in motion, with localised regions moving up (expanding) and down (contracting) with a maximum speed of ~500 m s^{-1}. In the following decade it was realised that these motions could be understood as vibrations (or oscillations) affecting the whole Sun (e.g. Figure 5.2).

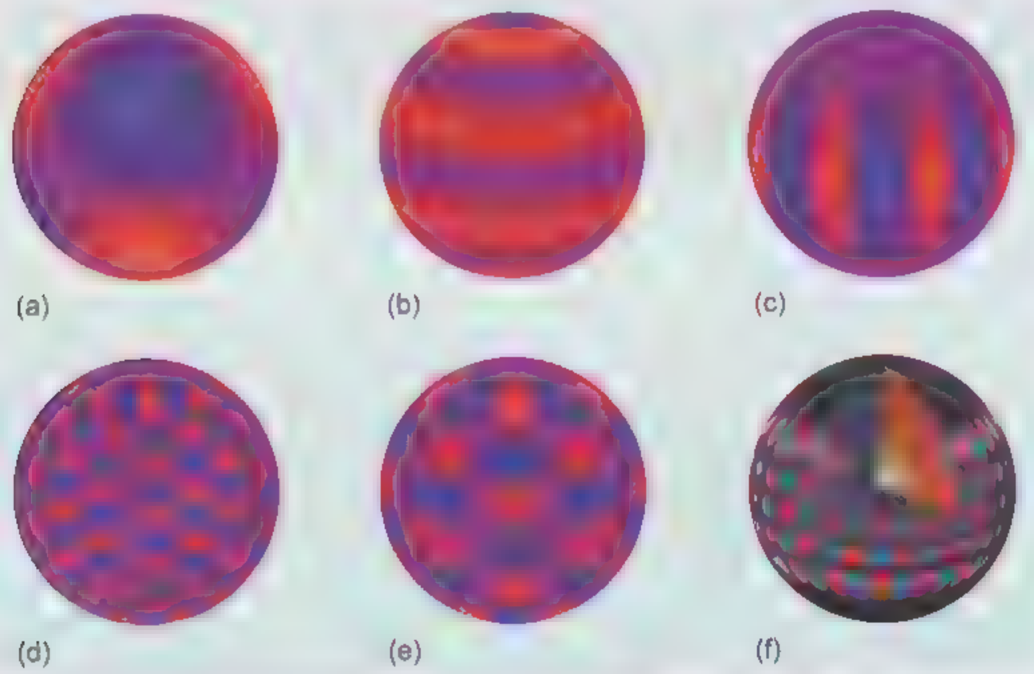


Figure 5.2 Global oscillations in the Sun. The red areas represent zones of temporary expansion, and blue areas of temporary contraction. Parts (a) to (f) show the motion of the solar surface for various different vibrations. Part (f) also shows a cutaway of the Sun exposing the internal motion of the Sun for one of these vibrations.

A drum will sound different if you put something inside it, e.g. a cushion. A pot in the kitchen will resonate when it's tapped, with the quality of the sound depending on its contents. Similarly, the vibrations of the Sun depend on the inner properties of the Sun such as temperature and density. The Sun's interior affects how sound waves pass through it, so by making accurate measurements of the motion of the surface of the Sun it's possible to determine the inner structure of the Sun. This is similar to how geologists have deduced the inner structure of the Earth, by measuring how the vibrations caused by earthquakes pass through the Earth.

5.2 The Sun as a star

Stars come in a wide range of temperatures and sizes (Section 1.5), with our own star the Sun being fairly middle-of-the-road – not the biggest nor smallest, nor the hottest or coolest. The combination of a wide range of different temperatures and sizes results in the amount of energy released by stars varying tremendously For example, the brightest stars we know of emit more than a million times more energy than our own Sun, while the faintest stars are thought to be a factor of ten thousand times fainter than the Sun.

You can see this in the young compact clusters of stars, such as the Jewel Box that you met earlier in Chapter 2.

- What is the difference between clusters of stars and constellations?

  The terms 'constellations' and 'stellar clusters' do not refer to the same objects. Constellations are patterns formed by the projections of stars onto the night sky that *in general* are not physically connected. One can imagine two stars appearing close together on the sky in the same constellation but one star being nearby and the second at a large distance. On the other hand, clusters consist of groups of stars that are in close proximity and are held together by their mutual gravitational attraction.

Clusters are perfect places to study how stars work since all the stars in a cluster must have formed at the same time and will also all have formed from the same material. The stars must therefore have had the same initial chemical composition (a topic which you will meet again in the next chapter). So, if the spectra of two stars indicate that their chemical compositions differ, this must be a result of the physical processes inside the stars, rather than having been formed from material with differing initial compositions.

The size of any cluster is much smaller than its distance from the Earth. Therefore, as seen from the Earth, all the stars within an individual cluster are at approximately the same distance. If two stars within an individual star cluster appear to differ in brightness, this must mean that the *total* light put out by each star must be different, rather than one star simply appearing fainter than another because it is further away. This *total* output of light is called the **luminosity**.

These facts make clusters very valuable for astronomers trying to understand the life cycle of stars, and the physical processes that drive them.

As Herschel remarked, the most striking properties of the stars in the Jewel Box (Figure 2.6) are their different brightnesses and colours. Given that they are all at the same distance from the Earth, this first observation confirms that some stars have a higher luminosity than others. Moreover, as you learnt in

Chapter 2, the differing colours of the stars imply differing temperatures. If all the stars in this cluster formed at the same time, why do they appear so different? This is the subject of this chapter.

5.3 Patterns in the stars: the HR diagram

Chapter 2 described how the spectra of stars could provide information on both the temperature and also chemical composition of stars. With a measurement of their distance (using, for example, one of the techniques described in Chapter 2) it's also possible to calculate their intrinsic luminosity – in other words, the total amount of light they radiate – from the amount of light received from them on Earth.

So, despite the immense distances involved, there's a lot of information available to astronomers. But how can all this detail shed light on the internal processes of stars? A common trick of scientists is to look for patterns of behaviour that give clues to the underlying rules of nature that are involved.

One of the simplest graphs of the properties of stars is a plot of their temperatures against their luminosities. This was first attempted by the Danish astronomer Ejnar Hertzsprung in 1906 and, independently, by the American Henry Norris Russell, using a more extensive set of data, in 1913. This single, simple diagram is arguably one of the most important in astronomy. It's the metaphorical key to unlocking the secrets of stellar evolution and today bears their names: the **Hertzsprung–Russell (HR) diagram**.

But what is so special about the HR diagram (Figure 5.3)? The answer is simply that it demonstrates that nature apparently only allows stars to have certain combinations of temperature and luminosity. There are also limits to the ranges of luminosity and temperature allowed for stars; the laws of physics appear to prevent the formation of very luminous and very faint stars. Both these observations provide clues to the inner workings of stars, so let's examine the HR diagram in a bit more detail.

The clearest feature to notice is that most stars occupy a thin strip in the diagram from top left (very luminous stars of high temperature) to bottom right (very faint, cool stars). In colour terms the most luminous stars known are bright blue, becoming redder as they become fainter. The Sun is on this strip too, as are 90% of stars. Since the vast majority of stars are found on this sequence it is known as the **main sequence**.

Above the cooler reaches of the **main sequence** are the **red giants**. These are cooler than the Sun, so they have a red-orange tinge to the visible light they emit. Living up to their names, the largest red giants may be over 100 times the size of our own Sun.

Stretching right across the upper regions of the HR diagram are the **supergiants** that cover a wide range of temperatures. This rarest subset of stars are true stellar behemoths. Placed in the centre of our Solar System the largest of these, such as the red supergiant VY CMa, would engulf all the planets out to the orbit of Saturn.

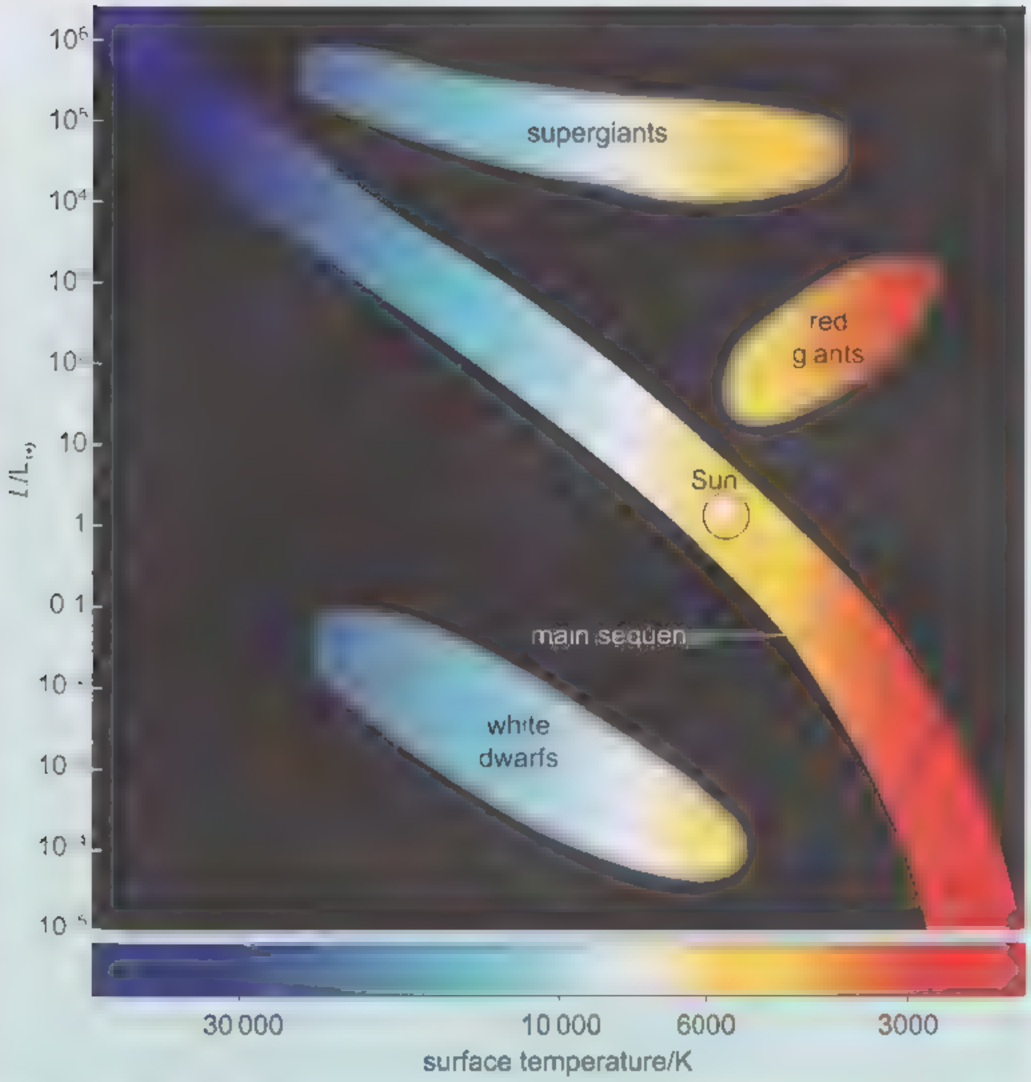


Figure 5.3 A Hertzsprung–Russell (HR) diagram. The vertical axis is luminosity, and the horizontal axis is temperature. Luminosities are measured as a proportion of the luminosity of the Sun, so $L/L_{\odot} = 10$ means the star is ten times more luminous than the Sun.

Finally, there is a group of very faint but hot stars in the bottom left of the HR diagram. These are called **white dwarfs** and are so faint that none are visible to the naked eye; their story forms part of the next chapter.

So how can this diagram be interpreted? There are two main facets to explain:

- why stars only occupy particular regions of the HR diagram, and
- why different numbers of stars are found in different regions.

Let's look at the second facet of the problem first. It could be explained by assuming that rather than remaining constant for all time, stars in fact change sizes and temperatures over their lifetimes. Stars will mostly be found where they spend most of their time, so stars must spend the majority of their lifetime on the main sequence. On the other hand, there are relatively few stars in the regions populated by red giants, so the proportion of their life cycle they spend there must be very much shorter than their main-sequence lifetime.

This might at first seem counter-intuitive after all our Sun hasn't changed over the course of recorded human history. However all this proves is that any changes are likely to occur over very long periods of time compared to a human lifetime millions or billions of years in fact. But what reason is there to believe that stars can only live for a fixed period of time? One good reason is that stars are continually emitting huge amounts of radiation and that some physical process must power this. Just as a burning candle consumes wax and eventually goes out, stars only have a finite supply of fuel and as this runs out their properties begin to change. At this point they move away from the main sequence. The next chapter will discuss what happens as stars run out of fuel.

Another reason to suppose that stars *do* change their properties is that some types of star, such as supergiants and white dwarfs, are missing in very young star clusters and are only found in older clusters. This suggests that certain types of star are not born with the properties that they have today, but instead have evolved into their present-day state.

The main sequence also has far more cool and faint stars than hot and luminous stars. It might also be reasonable to guess that the cool and faint stars live much longer than the hot, luminous stars. This turns out to be true, though the imbalance is also caused by different birth rates of bright and faint stars.

If you look carefully at how luminosity and temperature are represented on the axes of the HR diagram in Figure 5.3 you may find that they look slightly different from those of other graphs you have encountered before. Take a close look at the vertical axis. You will see that each division or notch marks an increase in luminosity by a factor of *ten* over the division immediately below. This is an example of a mathematical tool called a **logarithmic representation**. It is a way of handling and presenting very small and very large numbers simultaneously, like the idea of **exponents** you encountered in Chapter 2. There are many properties in astronomy that span very large ranges, such as the luminosities of stars, and distances between stars, galaxies and clusters of galaxies. Logarithmic representations are commonly used to make sense of these.

One example of a logarithmic scale that is used in astronomy and that you might have heard of is the **magnitude scale**, which is used to represent the brightness of objects such as stars and galaxies. One of the reasons astronomers use this system is due to an accident of evolution, since our eyes respond to changes in the brightness of objects in a logarithmic way. We perceive the *difference* in brightness between a 6 W and a 60 W light bulb as *identical* to that between a 10 W and 100 W bulb. Astronomers in previous centuries had to make estimates of the brightness of various stars solely by eye, so it was natural to make use of a logarithmic scale.

But this still doesn't explain why stars can't be found in every region of the HR diagram A clue to this is found in a physical property of stars that is not *explicitly* plotted in an HR diagram: their masses.

At this point you might reasonably ask how anyone can measure the masses of stars? The trick is to use pairs of stars orbiting each other, called binary stars. The stars are bound together by their mutual gravitational attraction. The stronger the gravitational force between them, the faster the stars orbit each other. Stars with bigger masses have stronger gravity, so they will orbit more quickly. (This is the same principle used to prove the presence of dark matter from the rotation of whole galaxies, as discussed in Section 3.2.)

In practice two more things are needed for this technique to work. Firstly, you need to be able to detect narrow features such as absorption lines in the spectra of both stars in order to be able to measure their speeds via the Doppler effect (Section 2.5.3). Secondly, you need to know the orientation. To see why, imagine you're looking at the stars from directly above or below their orbit, as shown in Figure 5.4 (point A). The motion of the stars would be side-to-side, so you wouldn't see any Doppler effect. Now imagine you're aligned with the orbit, as also shown in Figure 5.4 (point B) This time, the stars move towards and away from you, so you *do* see a Doppler effect. If the stars were orbiting each other at some other random, oblique angle to our line of sight, you'd see a weaker Doppler effect.

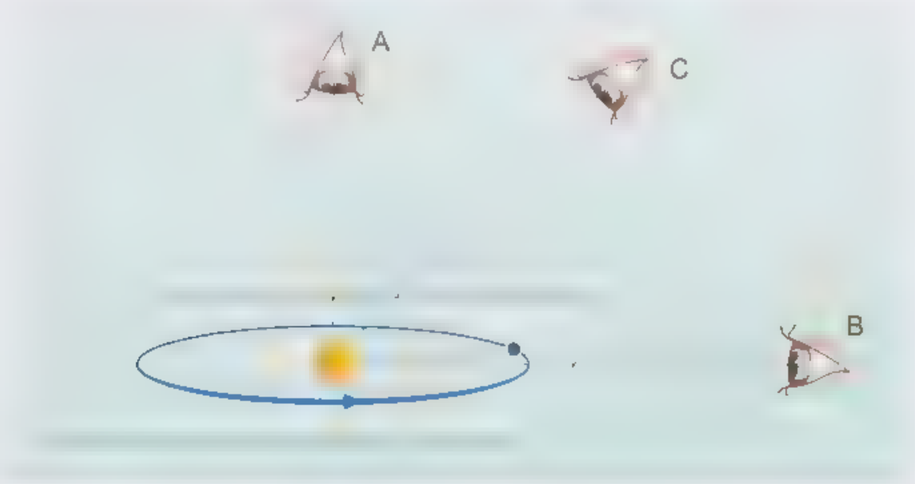


Figure 5.4 A binary star seen from above (point A), from the side (point B) and at some other random angle (point C). Only when the line of sight is aligned with the orbit (point B) can you see an eclipsing binary, and get reliable measurements of the masses of the stars.

The trick is to use stars that pass in front of and behind each other, known as eclipsing binary stars This causes clear changes in the brightnesses. These binary stars are the most useful for measuring masses from their orbital speeds.

- What would happen if you estimated the masses of stars in a binary system that's *not* eclipsing?

  The stars would be orbiting at some random angle to your line of sight, so the Doppler effect would be smaller. Therefore, you'd infer a speed that's too small, and so the masses you'd derive would be underestimates.

There's a startling trend of masses along the main sequence: the more massive a star is, the hotter and more luminous it is, and vice versa. This on its own rules out one explanation for the main sequence: that all stars on it are of the same mass and that it's an evolutionary sequence.

- Why is an evolutionary sequence ruled out?

  An evolutionary sequence would need stars to become progressively hotter and brighter (or cooler and fainter) as they age, while keeping the same mass. The fact that stars along the main sequence have very different masses means that the main sequence can't be an evolutionary sequence.

So the HR diagram not only separates stars on the basis of their luminosities and temperatures, but also on the basis of their masses, at least while they are located on the main sequence. Why should the mass of a star be directly related to its luminosity and temperature? The key to this is their structure and the physical process that powers them.

5.4 Energy generation in stars

Let's review what's been covered so far about stars. Firstly, they need a power source since they are constantly radiating energy away into space. If they didn't have one they would gradually cool and fade away until they were invisible. Secondly, the HR diagram implies that the more massive a star is, the more luminous it is as well. Therefore, more massive stars must produce more energy than their less massive siblings.

An obvious explanation for this behaviour would be that heavier stars are more luminous simply because they have more fuel to burn, which is why they produce more energy. To take a simple example, imagine a star that's ten times as heavy as the Sun. This must have ten times as much fuel available to it as our Sun. It would be ten times as bright as the Sun if it burned its fuel at the same rate. However, such stars are approximately *ten thousand* times brighter than the Sun!

In fact the luminosity of a star is very strongly dependent on its mass. On average, the luminosity of a star is proportional to its mass to the power four (or M^4 – that is its mass times its mass times its mass times its mass). This means that a star 2 times as heavy as the Sun would be $2 \times 2 \times 2 \times 2 = 16$ times as luminous, while a star half the mass would be only $0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0625$ times as bright.

This is quite an unexpected and profound finding, so it's worth thinking about what this means in a little more detail. Firstly, this relationship can be used to work out how long stars live. Imagine again a hypothetical star that is 2 times as heavy as the Sun. Since it is 16 times as bright as the Sun it must be burning its fuel 16 times more quickly. Unfortunately, it has only 2 times as much fuel available to it and so will run out of fuel much more quickly than the Sun. At that rate it will only live for 2/16 = 1/8 of the lifetime of the Sun.

■ How long would a star half as massive as the Sun be expected to live for?

□ The star would only be 0.0625 times as bright as the Sun (see above) but would have half as much fuel as the Sun, so it would live for 0.5/0.0625 = 8 times as long as the Sun.

So, a more massive a star has a shorter lifetime. This neatly explains the observations that that there are far fewer luminous stars than faint stars on the HR diagram Even if the number of high-mass and low-mass stars formed in a star cluster (or a galaxy) were the same, the longer lifetime of the low-mass stars would quickly result in a greater proportion being present as the more massive stars rapidly ran out of fuel.

This finding can also help reveal something about how stars generate their energy. To do this, the extra clue that's needed is the fact that stars on the main sequence seem to be rather stable. They don't appear to swell up or shrink quickly, because if they did that then life on Earth would have either baked or frozen. It's not unreasonable to suppose that the Sun's been largely unchanged for the billions of years that life has been on Earth.

What this means for the Sun – and indeed all other stars – is that the force exerted by gravity, which is trying to cause the Sun to contract under its own weight, must be exactly balanced by forces directed outwards from its core which support it against its own weight. If gravity were to win out the Sun would rapidly collapse in on itself. On the other hand, if the supporting forces won out then the Sun would explode outwards. If either extreme were true then the Sun wouldn't be as stable as it is. This balance of forces is known as **hydrostatic equilibrium**.

Understanding hydrostatic equilibrium and the detailed internal structure of a star involves complicated mathematics developed by Sir Arthur Eddington and others. Their analysis included the way that gases heat up when they're compressed. (This connection between temperature and pressure is why a bicycle pump heats up when you use it. It's also the reason you should never heat up a can of food without removing the lid or piercing the tin to release the pressure created as you heat it.) The gas in the centre of a star is being crushed by all the gas above it, so this is an important consideration. Without going into the mathematical details, a consequence of hydrostatic equilibrium turns out to be that the interior temperature of any star is proportional to the star's mass, so twice the mass means it's twice as hot in the interior.

This is the critical difference between high- and low-mass stars: the cores of high-mass stars are hotter than the cores of low-mass stars. The HR diagram might make this seem a rather obvious statement to make, but remember that this is a plot of the *surface* temperature of the star. There's no way of *directly* measuring the internal core temperature of any star because no-one can see the light that it emits. The only light that comes out of a star is the light that has escaped after a tortuous journey through the cooler outer layers of the star, and which has been modified by its passage.

Nevertheless this is the key to understanding the relationship between the increase in stellar luminosity with mass The rate at which energy is released in a star – and hence its luminosity – appears to be directly related to its internal (core) temperature, which *in turn* depends on the star's mass. But why does the energy release depend on the core temperature? This can be answered by examining the nature of the nuclear reactions that power stars.

5.5 The nuclear reactions that power stars

As far back as 1917 Sir Arthur Eddington had suggested that the energy source of stars was sub-atomic in origin, and involved the conversion of hydrogen into other elements and the simultaneous release of energy. But how does this process work in detail? At its heart it involves Einstein's famous equation:

$$E = mc^2.$$

This deceptively simple equation essentially shows the equivalence of energy (E) and mass (m), and raises the possibility that matter may be directly converted to energy (and vice versa).

The final piece of the equation is c^2, where c is the speed of light Now c is a very large number (3×10^8 m s^{-1}) and it indicates just how much energy is released when matter is annihilated. To illustrate this let's imagine making a cup of tea A single sugar cube converted completely to energy would release enough to power a normal kettle to run continuously for about 10 000 years!

- If you fully convert 1 kg of mass to energy, how long would this allow a 60 W light bulb to shine for?

  The amount of energy released would be 1 kg × (3×10^8 m s^{-1})^2 = 9×10^{16} kg m^2 s^{-2} or 9×10^{16} joules (where the joule (J) is a unit of energy which has units of kg m^2 s^{-2}). Now 1 watt = 1 J s^{-1}, so a 60 W light bulb would run for 9×10^{16} J/60 J s^{-1} = 1.5×10^{15} s or approximately 47.5 million years!

This astonishing equivalence of vast quantities of energy to small quantities of mass is the reason nuclear reactions can keep stars shining for billions of years. Every second, the Sun emits 20 million times more energy than the USA generates each year from fossil fuels, and to do so it must convert about 4.3 billion kg of its mass to energy. But the Sun's mass is 2×10^{30} kg, so in practice it only loses 2×10^{-21} of its mass every second. At this rate it would take about 15 thousand billion years to run out of fuel if *all* the mass of the Sun were converted to energy! In practice however, not all the mass of the Sun can be converted, and so its lifetime is expected to be 'only' 10 billion years or so.

Nevertheless, this calculation demonstrates why nuclear reactions can serve as the energy sources of stars. But what are these reactions and how do they work? In essence, nuclear reactions involve combining or splitting atomic nuclei to form new nuclei of different mass. This is the crucial difference between chemical and nuclear reactions: chemical reactions do not affect the nuclei of atoms involved with them.

As you saw in Section 4.8 an atom consists of a massive, compact positively charged nucleus with one or more negatively charged electrons around it. The nuclei of atoms are composed of two different types of particle of almost identical masses: positively charged protons and electrically neutral neutrons. Different elements have differing numbers of protons present. Hydrogen, for example has one proton, helium has two, carbon has six, oxygen has eight and so on. The number of neutrons present doesn't affect the identity of an atomic nucleus, so an atom of hydrogen can contain zero, one or two neutrons and still be hydrogen. Chemically, these three **isotopes** are identical to one another.

Now the key – and counter-intuitive – fact that makes nuclear reactions the source of the Sun's energy is that the masses of nuclei can't be calculated by simply adding up the number of individual protons and neutrons that are present. As an example let's consider the masses of isotopes of hydrogen and helium. The simplest possible isotope of hydrogen comprises a nucleus containing just one proton. Now the physical processes that release energy in the Sun are a series of **nuclear fusion** reactions that ultimately fuse four such hydrogen atoms together to produce one atom of helium, consisting of two protons and two neutrons (some protons are converted to neutrons in the process). Since the masses of protons and neutrons are almost identical, you might expect that the total masses of the four protons equals that of the single helium nucleus. However, the mass of the helium nucleus is actually slightly *smaller* than the mass of the four protons! So somewhere in the reaction we've lost mass. This missing mass has been converted to energy, in accordance with Einstein's equation. It's this conversion of mass to energy which powers the Sun and all other stars.

The difference in mass is very small. The total mass of four protons is 6.692×10^{-27} kg and that of a single helium nucleus is 6.645×10^{-27} kg, so the energy released in this reaction is tiny. Powering the Sun must need a *lot* of reactions – in fact a staggering 8.7×10^{37} reactions are needed every second! A lot of hydrogen must therefore be present to fuel the Sun. But the Sun's mass is 2.0×10^{30} kg, which turns out to be more than enough.

The conversion of hydrogen to helium is termed **hydrogen burning**, although this is nothing to do with setting anything on fire! Fire is a chemical reaction, not a nuclear reaction. The main nuclear reactions in hydrogen burning are known as the proton–proton chain.

The stages of the proton–proton chain are beyond the scope of this book, but it's helpful to examine the first stage because it illustrates a very important point. This is the fusing of two hydrogen nuclei (protons) together. The two nuclei involved in the reaction have identical positive charges, so they repel each other with greater and greater force the closer they get to each other.

This presents a problem, because the protons need to be close to one another in order to interact and undergo fusion. In practice this means that both particles need to be moving very rapidly towards one another in order to overcome this repulsion. Now the higher the temperature of a gas, the faster the atoms or molecules of that gas are moving inside it. This is how stars with higher internal temperatures produce more energy: the particles are moving more quickly, so they find it easier to overcome their mutual repulsion, so they can interact more readily.

The higher core temperatures of stars more massive than the Sun also allow other nuclear reactions to occur that involve nuclei of heavier elements: carbon, nitrogen and oxygen (in a process called the CNO cycle). Although the net effect is still to convert hydrogen to helium, these heavier nuclei have stronger electric charges, so higher energies and temperatures are required to overcome their much greater mutual repulsion As a result, it turns out that the energy released by both these reaction chains is *astonishingly* sensitive to temperature. The rate of energy released by the proton–proton chain is proportional to the temperature of the stellar core to the power of four (i.e T^4 or $T \times T \times T \times T$) while for the CNO cycle it is proportional to the T^{17}, i.e.

$$T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T \times T.$$

For example, if you just double the core temperature, the energy released by the CNO cycle rises by a factor of $2^{17} \approx 130\ 000$. This is the underlying reason why the luminosity of stars rapidly rises with stellar mass: more massive stars have hotter cores, while nuclear reactions have extreme sensitivity to temperature, so much more energy is released by the nuclear reactions in more massive stars.

Is there any *direct* observational evidence that this process is indeed occurring in the Sun? It turns out there is, but it has proved very difficult to come by since it requires detecting a ghostly by-product of nuclear fusion (briefly mentioned previously in Chapter 4): the **neutrino**. Neutrinos are particles with very low mass and no electric charge. Because of this they do not interact electrically as protons and electrons do, so they're almost completely unaffected by normal matter. Indeed, you'd need a block of lead over 1 light-year in depth in order to stop a neutrino from the Sun in its tracks! But precisely because of this, they can escape from the core of the Sun, directly carrying information on the processes occurring there, if only one could detect them Technically, this has proved very difficult; efforts began in 1970, but it took over 30 years for an experiment to detect them and to confirm the predictions of theories of the deep interior of the Sun.

5.6 Stellar masses

The nuclear processes powering stars also help explain why stars only occupy a comparatively small range of masses. At very low masses the core of the would-be star never becomes hot enough to permit hydrogen nuclei to overcome their mutual repulsion and undergo fusion. Astronomers currently believe the lower limit to the mass of stars is around $0.08\,\mathrm{M}_{\odot}$.

At the other extreme, the energy released by stars rises very rapidly with increasing stellar mass. As this happens the radiation released by the process of nuclear fusion imparts an increasing pressure on the outer layers of the star. This pressure has to be balanced by gravity for the star to remain stable. But energy production increases very quickly as you choose stars with larger and larger stellar masses and this increase greatly outstrips the increase in the force of gravity. Eventually, at some point the pressure from radiation wins out and gravity can no longer hold the star together. This means there must be a limit to the mass of stars above which they cannot remain stable. The precise location of this limit is currently uncertain but is expected to be above $100\,M_{\odot}$. Observations of very hot and luminous stars located in regions such as the Arches cluster in the centre of our galaxy in the next few years will help us to determine this upper limit more accurately (Figure 5.5).

Figure 5.5 The Arches cluster, located in the central region of our galaxy. It is thought to host some of the most massive stars allowed by the laws of physics, with mass over 100M

In summary, mass is the key to understanding the position of stars within the HR diagram. Stars of different masses occupy different regions of the main sequence, which, contrary to appearances, does not correspond to an evolutionary track. An obvious question then arises – if the main sequence is not populated by identical stars at different stages of their lives, then where are the different stages of a star's life? What happens when stars run out of hydrogen? Can it then make use of other elements as fuel? If so, how does this affect the structure? In particular, do its temperature and luminosity evolve? Such changes would move a star's position on the HR diagram. These points will be addressed in the next chapter.

End-of-chapter questions

Question 5.1 Section 5.1 contains a prediction of how long a star half as massive as the Sun would live for. Why have astronomers not yet been able to test this prediction with observations?

Question 5.2 Astronomers can probe the process of nuclear fusion in the cores of stars by studying the neutrinos released in these reactions. Considering a star twice as large as the Sun, would you expect it to emit fewer, the same number, or more neutrinos than the Sun and why?

Question 5.3 A friend tells you about a television programme that claimed that astronomers had discovered a star 1000 times the mass of the Sun and twice as old. After reading this chapter, why might you conclude that someone is mistaken?

Now go to the module website and do the remaining activities associated with Chapter 5.

Activity 5 2 · 40mins
Activity 5 3 · 30mins

Chapter 6 The cosmic cycle: the death and birth of stars

6.1 Introduction

The previous chapter covered the observational properties of stars on the main sequence and investigated the physical processes that cause them to shine. This showed that stars cover a large but finite range of luminosities, temperatures and masses. Moreover, the nuclear processes which power stars depend on the core temperature – and hence stellar mass – although in *all* stars energy is released by the fusing together of light nuclei to form heavier elements. The process of nucleosynthesis was not just restricted to the early Universe, but has continued ever since the first star was born.

However, Chapter 5 also showed that stars only have a fixed amount of fuel available to them and so a limited lifetime. The very smallest stars burn their fuel so slowly that their lifetime is expected to be over a trillion years, which is over 70 times longer than the current age of the Universe! On the other hand, the most massive are thought to live for only about 3 million years. Nevertheless, eventually every individual star has to die and in this chapter we will investigate how this occurs, what remnants dead and dying stars leave behind and how this in turn leads to the next generation of star birth from the ashes of their forebears.

This cyclic process and the contents of this chapter are beautifully illustrated by the two images of the interacting Antennae galaxies shown in Figure 6.1. In the optical image (Figure 6.1a) there are brilliant bursts of new star formation, with each of the blue-white blobs consisting of a massive stellar cluster containing hundreds of thousands of newborn stars (their blue colour being due to their high temperatures). Surrounding this there are remnants of the clouds of gas from which they formed (glowing in the red pink light characteristic of hydrogen) separated by dark lanes produced by dense clouds of gas and dust within which lies the raw material for the next generation of stars. And yet the same regions, when observed in X-rays (Figure 6.1b), reveal the corpses of the preceding generation of stars, either radiating their residual heat away into space as they quietly cool over millions or billions of years or actively generating their own energy as they tear material from companion stars.

6.2 Life after the main sequence in Sun-like stars

To the best of our knowledge the Sun is approximately 4.6 billion years old and is likely to live for roughly as long again. Considering the quantity of hydrogen available to it this might seem like a comparatively small number. Indeed, if *all* the hydrogen within the Sun were converted to helium, enough energy would be released to keep the Sun shining at its current luminosity for about 100 billion years.

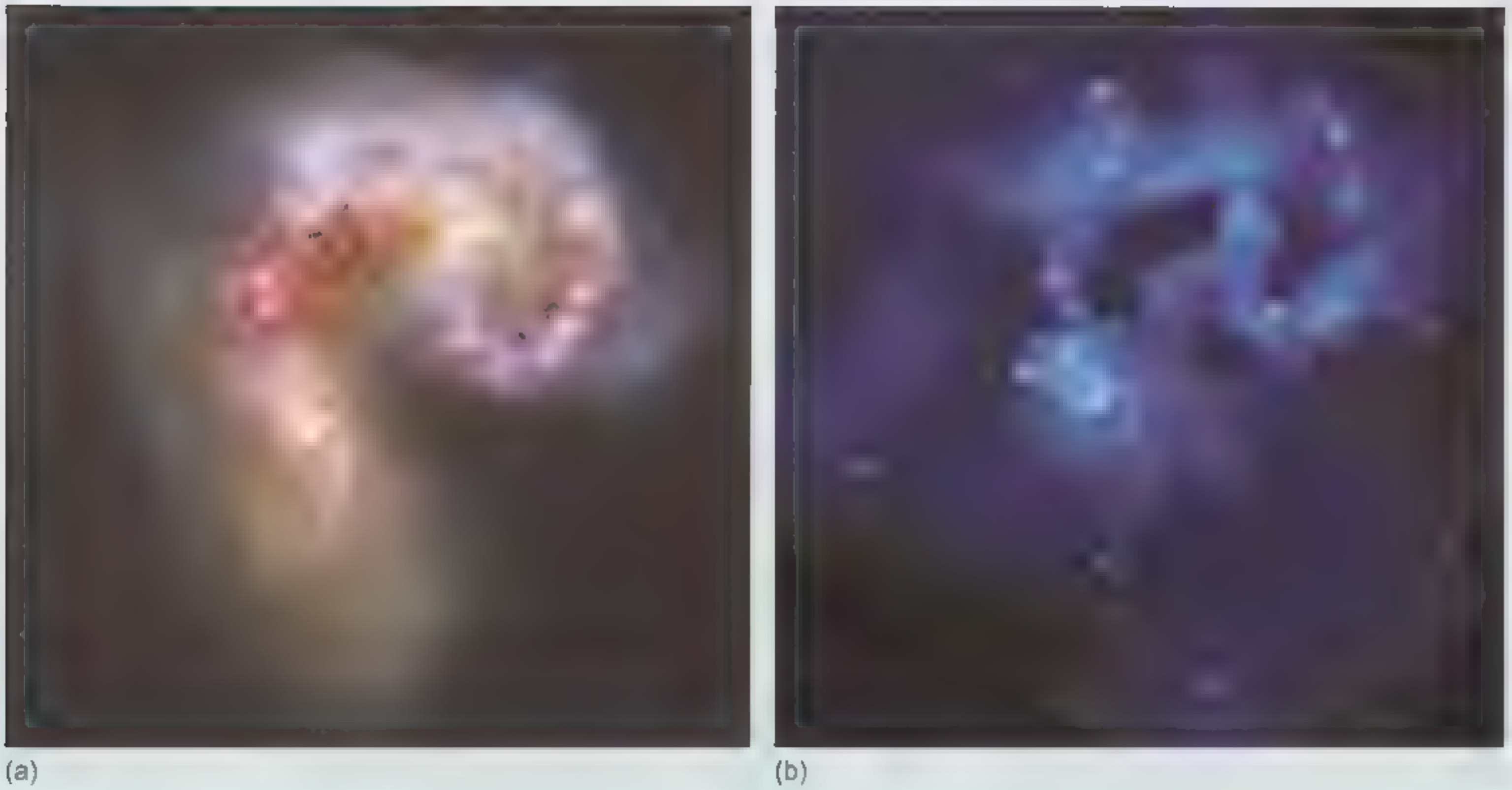


Figure 6.1 (a) Optical and (b) X-ray images of the interacting star-forming Antennae galaxies, demonstrating regions of current star formation, clouds of cold dusty gas waiting to form stars (optical image) and the dead remnants of the preceding generation of stars (bright point sources in the X-ray image).

■ Why is this lifetime much shorter than the potential lifetime of 15 thousand billion years given in Section 5.5?

Because not all the Sun is made of hydrogen, and not all the mass can be converted into energy. It's only the difference in mass between the four hydrogen atoms and the helium atom that they produce that's converted into energy.

Astronomers believe that this discrepancy is due to the fact that the pristine hydrogen in the outer layers of the Sun can't easily be transported into the core, where temperatures are high enough for fusion to occur. This is due to the fact that the convective zone – in which gas physically moves and which might therefore transport hydrogen to fuel the nuclear reactions – does not extend to the core, being separated from it by the radiative zone, where it is thought that there are no large-scale flows of matter (Figure 5.1) Consequently, the only fuel that is available to the core is that with which it was born. This comprises roughly 10% of the total mass of Sun, explaining the difference between the Sun's potential lifespan and the greatest age it is likely to achieve.

So what happens when a star like the Sun runs out of fuel? This will be a *gradual* process, but as it occurs the rate of energy production will also decrease. This in turn breaks the delicate balancing act between the outward pressure of the hot gas and the inward force of gravity that the star has performed up to then. The core will then begin to collapse under its own weight, because the force due to gravity will become larger than the thermal pressure due to nuclear fusion. As it collapses, the core will begin to heat up as gravitational potential energy is converted to thermal energy.

One of the most basic assumptions in physics is that in any physical process energy cannot be created or destroyed, merely converted from one form to another. This is known as the **conservation of energy** and you've already seen this in mathematical form via Einstein's famous equation $E = mc^2$ where energy in the form of mass (m) can be converted to some other form. We all experience the conservation of energy in our everyday lives. For example, light bulbs turn electrical energy into heat and light energy. Likewise, chemical energy is converted all the time by our muscle cells into kinetic (or movement) energy. Chapter 8 will talk about energy in life in more detail.

A more abstract concept is that of **gravitational potential energy**, which represents the energy stored in an object by virtue of its position in a gravitational field. The best way to understand this is to imagine picking up a mug from a table and placing it overhead on a shelf. In this action chemical energy has been converted to kinetic energy in your muscles, which leads to your arm and cup moving. However, you've had to work against gravity to move both arm and mug and consequently some of the energy from your muscles is now 'stored' in the mug. You can see energy is stored because if you were to accidentally knock the mug off the shelf it would fall and then smash when it hits the floor – converting gravitational potential energy first to kinetic energy as it falls and then heat and sound energy when it hits the floor.

This concept of objects converting potential energy to heat and kinetic energy as gravity acts upon them is fundamental to many fields of astronomy. You'll be encountering this concept at several points in this chapter.

As the core contracts it also heats up the layers above it, which eventually become hot enough to permit nuclear fusion to occur in an overlying shell of unprocessed hydrogen in exactly the same way it previously did in the core. This is known as **hydrogen shell burning**. While this is occurring, the core continues to contract and heat up as gravitational potential energy is converted into thermal energy. Eventually, the core temperature reaches 10^8 K, at which point a new range of nuclear reactions becomes possible: the fusion of helium into heavier elements. The net effect of the reaction is the conversion of three helium nuclei into a carbon nucleus. This is called **helium burning** and the reaction chain is the **triple-alpha process**, since α (alpha) is the commonly used symbol for a helium nucleus ${}^{4}_{2}\text{He}$.

The triple-alpha process releases energy, which in turn stops the contraction of the core and stabilises the star. However, the process releases only a few percent of the energy per kilogram of fuel compared to the energy released by the fusion of hydrogen fuel to helium. This means that the star must consume its fuel at a much higher rate in order to support itself against gravity, and hence this portion of its life is correspondingly shorter than the main sequence hydrogen-burning phase.

While the core of the star contracts, triggering first hydrogen shell burning and then helium core burning, what happens to the surface properties? A full explanation involves the mathematical equations of stellar structure. These are well beyond the scope of this short module, but the resulting effects are straightforward to describe. As hydrogen shell burning progresses, the rate of energy production rises. The energy that's released is carried to the surface via convection, as are the products of the nuclear burning. At the same time the radius of the star increases, so its surface area also increases. There's now a bigger supply of energy, but it's still not enough to keep the outer layers of the star at their previous temperature, so the temperature of the photosphere decreases to below 4000 K. This corresponds to an orange–red colour. The star has now become a **red giant**, and it has moved on the HR diagram to regions of higher luminosity but cooler temperature (Figure 6.2). This process is called ascending the **red giant branch**.

It is thought that for a very brief period of their lifetime white dwarfs may evolve to very high surface temperatures before cooling and fading.

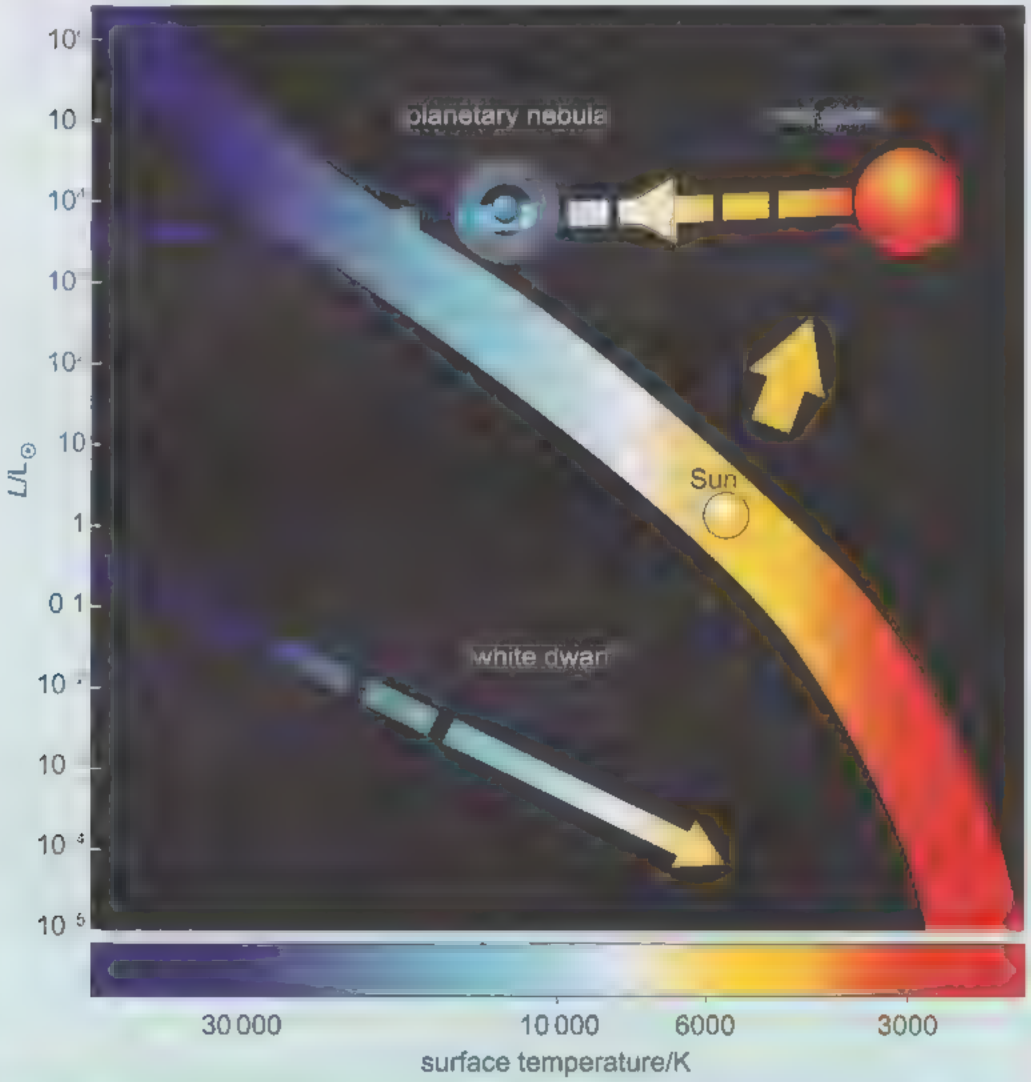


Figure 6.2 The HR diagram, with evolutionary track of a Sun-like star.

As you may remember from the last chapter, the region of the HR diagram occupied by red giant stars is more sparsely populated than the place on the main sequence where they came from. This is because of the much shorter lifetime of this phase, and you'll now appreciate that this is due in turn to the relative inefficiency of helium burning compared to hydrogen burning. The star has responded to the exhaustion of its original hydrogen fuel by starting the fusion of helium into heavier elements. This continues the process of nucleosynthesis, fusing lighter elements to make heavier ones.

There's another interesting reaction at this stage: each carbon nucleus in the core can capture a further helium nucleus to make an oxygen nucleus. This process and the triple-alpha process are thought to be the main sources of oxygen and carbon in the Universe.

6.3 The final stages in the life of Sun-like stars

Does this process of fusing light elements to make heavier ones continue indefinitely? The answer is emphatically no. Once the core runs out of helium, it contracts again under its own gravity. The contraction heats the core up. The temperature rises so much that helium burning starts in a shell around the core. The star now consists of an inert core made of carbon and oxygen, surrounded by concentric helium-burning and hydrogen-burning shells. This causes a further expansion and cooling of the photosphere, at which point the star's radius will be comparable to the radius of the orbit of Mars!

However, helium shell burning is not a stable process and leads to a series of **thermal pulses**, involving expansion and contraction of the shell, and a corresponding variation in energy output. A pulse might only last a few thousand years. During these events further nuclear reactions can occur in the shells, which may lead to the production of elements as heavy as bismuth (atomic number 83, i.e. containing 83 protons). These reactions take a relatively large amount of time, so astronomers have called them **s-process reactions**, where the 's' simply stands for slow.

The thermal pulses also alter the pattern of convection within the stars, which transports the newly synthesised elements, for example technetium (atomic number 43), to the stellar surface. Technetium is unstable and quickly decays to form lighter elements. Nevertheless, technetium is seen at the surface of stars, which means it must be continually replenished at the surface. This is exactly what's predicted from convection, so this is an important confirmation of astronomers' theories of evolving stars. If technetium were not continually being forged within the star via the s-process and brought to the surface via convection, astronomers would not see its steady signature in stellar spectra.

The thermal pulses can be so strong they even push off the outer layers of the star in a series of dramatic events. This is thought to be the origin of **planetary nebulae** (e.g. Figure 6.3), beautiful shells of gas ejected from and heated by the central dying star. Be warned: the name is an unfortunate historical accident. When they were first discovered and observed through small telescopes, they looked like gas giant planets, which gave rise to their name. However, they are not directly connected to planets or to planet formation. The only link is that they have taken heavy elements produced by nucleosynthesis within the star and injected them back into space, and these heavy elements are raw materials from which planets can one day form. Planetary nebulae come in a bewildering range of shapes and sizes. The reasons for this are currently uncertain, but rapid rotation of the star and interaction with a binary companion (or a combination of both) have both been suggested to explain this behaviour.

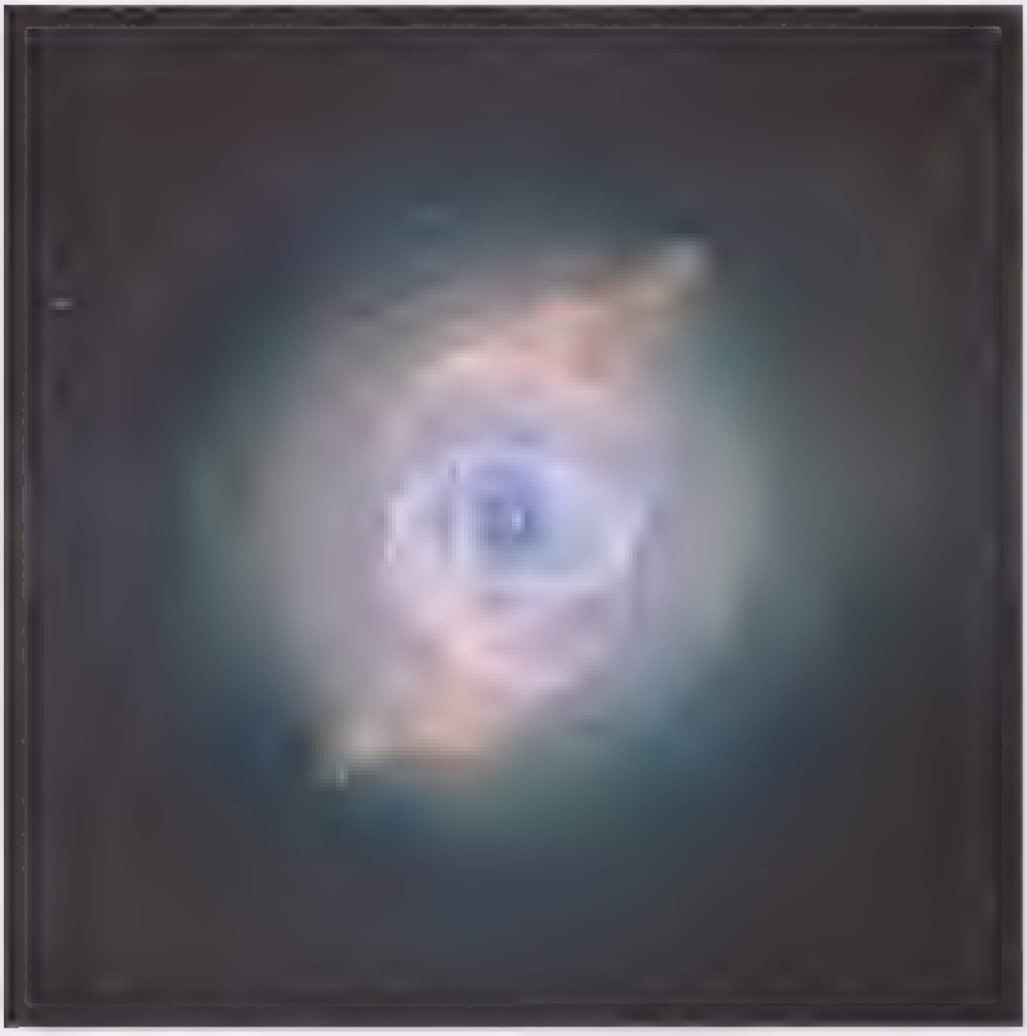


Figure 6.3 The Cat's Eye planetary nebula with its central white dwarf.

In the absence of an energy source, the inert carbon–oxygen core of the star continues to contract and heat up. Can it repeat its previous trick and reach temperatures high enough to produce energy via the nuclear burning of one or both elements? For stars of up to approximately eight times the mass of the Sun the answer is thought to be no. The reason is that densities in the core become so high that another force becomes important, and this new force halts the contraction before temperatures become high enough for carbon-burning or oxygen-burning reactions to be possible. Unlike the thermal gas pressure that has supported the star so far, this force is a peculiar result of the quantum behaviour of electrons. It can be understood as the inability of two identical particles – electrons in this case – to occupy the same space at the same time. This in turn leads to a pressure when electrons are pressed too closely together called **electron degeneracy pressure**. This force arises from the close proximity of electrons to each other, so it depends on the density of the core rather than its temperature, but nevertheless it supports the core against further gravitational collapse The carbon oxygen core will eventually be revealed as a white dwarf star about which you will learn more in Section 6.6.

6.4 The post-main sequence life of massive stars

The last section described the life cycle of lower-mass stars, up to about eight times the mass of our own Sun. However, stars can be 100 times the mass of our Sun, or more. How do those stars end their lives? Let's start by re-examining the fate of lower-mass stars. Once a star runs out of fuel in its core, the core contracts. The overlying material then falls in and heats up until nuclear reactions start in a shell around the core. While this happens the core continues to contract and heat up until it reaches a point when it's hot enough to burn the nuclear 'ash' of the preceding fusion reactions. This in turn carries on until the supply of fuel for the new set of reactions is exhausted at which point the core contracts again until the region surrounding it can maintain fusion reactions.

The end result is a nested shell structure within the star, with the increasing temperatures encountered as you move towards the core supporting the fusion of progressively more massive nuclei.

- Why does the fusion of more massive nuclei need higher temperatures?

  More massive nuclei have a larger number of protons, so they have a larger total positive electric charge. This results in greater mutual electrical repulsion, which the nuclei need to overcome in order to achieve nuclear fusion. They therefore need to be moving more quickly, which means they need to have a higher temperature.

Section 5.4 showed why the core temperatures of massive stars are systematically higher than those of their less massive relatives. The cores of stars with masses above about $8\,M_\odot$ reach temperatures of 5×10^8 K, at which point **carbon burning** starts. This fuses two carbon nuclei to make neon (with a new helium nucleus created in the process). Once the carbon runs out, the star's core contracts again, raising its temperature until **neon burning** at 1.5×10^9 K is possible This fuses a neon nucleus with a helium nucleus, to make magnesium. After the neon fuel runs out, a further bout of contraction raises the core temperature to about 2×10^9 K, at which point **oxygen burning** creates silicon Finally, the process of core contraction is repeated and **silicon burning** begins at core temperatures of 3×10^9 K, creating sulfur, argon and calcium. This sequence of nuclear reactions continues and ultimately leads to the production of a number of heavier elements in the so called **iron group** of elements, such as iron, chromium, manganese, cobalt and nickel. You can now see where some of the elements that make up the human body (Table 1.2) come from!

At the end of this sequence of nuclear reactions the star is like a 'cosmic onion'. It has a predominantly iron core at a temperature of about 7×10^9 K, surrounded by concentric shells of silicon and sulphur, oxygen and carbon, helium and possibly an outer layer of hydrogen, each at a lower temperature than the shells within.

What effects do these nuclear reactions have on the outwardly visible properties of such stars? As in low-mass stars, these extra sources of energy cause the star to expand and then cool, although their higher initial temperatures mean that they typically become hot, blue supergiants before evolving to even greater sizes and cooler temperatures, passing through both yellow and red supergiant phases (e.g. Figure 6.4 – compare with Figure 5.3). The largest known stars are these red supergiants. If one were placed at the centre of the Solar System, it would engulf Jupiter!

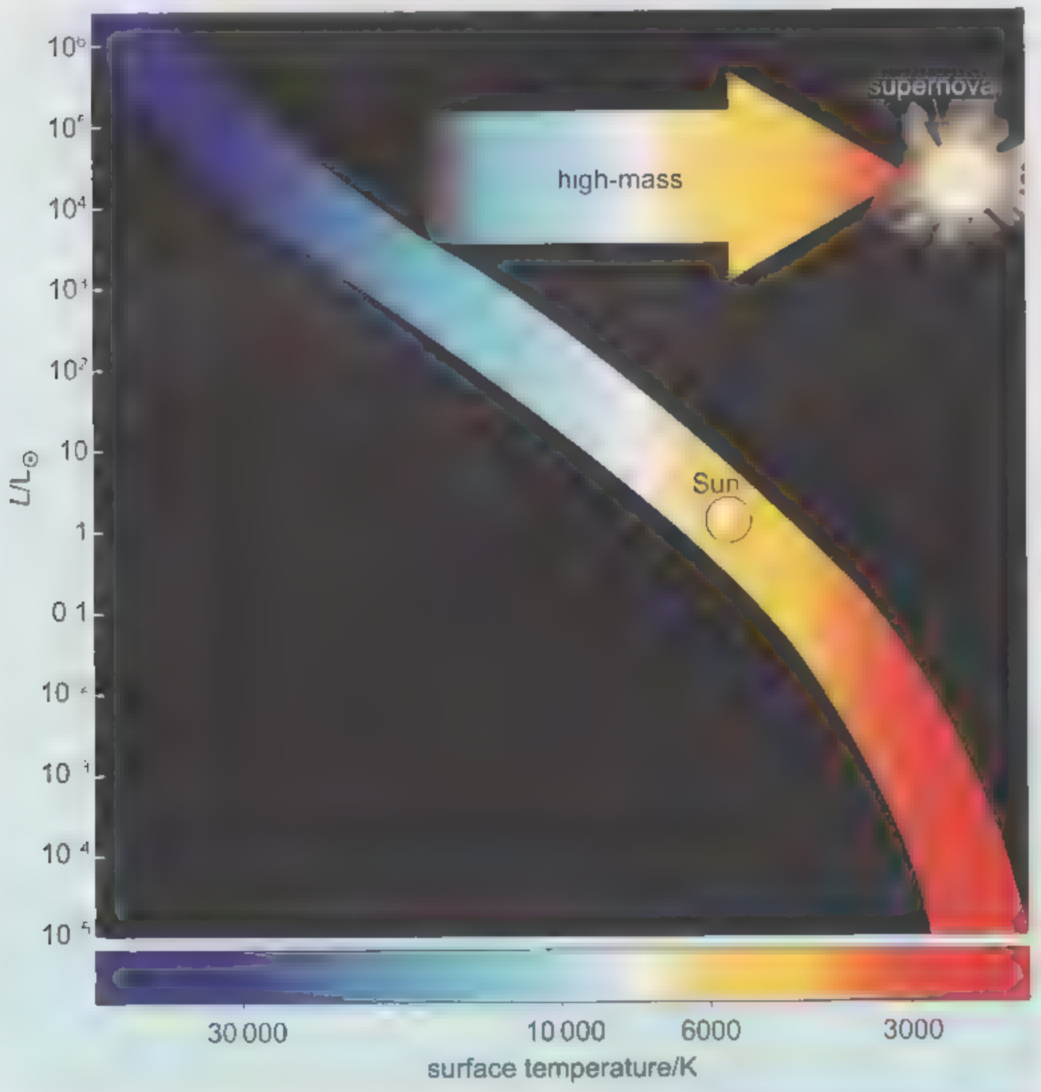


Figure 6.4 Evolutionary track of a high-mass star across the HR diagram.

Unlike Sun-like stars, astronomers are far less certain about the details of the post-main sequence life of very massive stars. The huge quantities of energy they release can be almost enough to overcome gravity. This makes the stars unstable, and the effects of this instability are difficult to predict. Broadly though, they are expected to be constantly ejecting large amounts of gas from their surfaces. This profoundly influences their evolution. A famous example is one of the most massive stars known in our galaxy, Eta Carinae. It's believed to have been born with a mass well over a hundred times that of the Sun and it's now more than four million times brighter than the Sun! In the nineteenth century it underwent an explosive outburst and became even more luminous. In the process it formed a compact nebula around itself (Figure 6.5). This nebula appears to have formed in only 30 years yet the amount of gas it contains is at least 15 times the mass of our Sun!

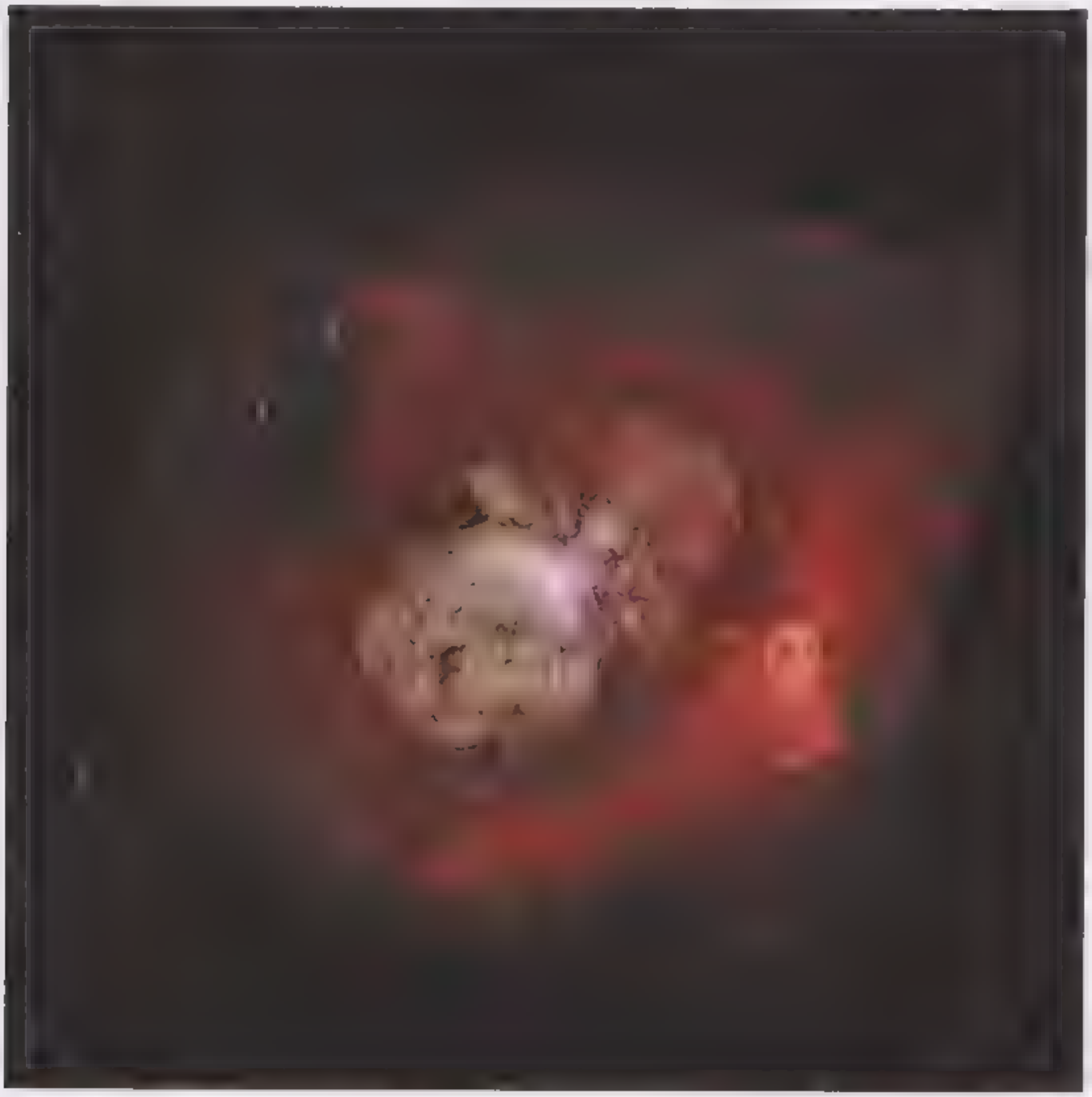


Figure 6.5 Eta Carinae, one of the most massive stars known, surrounded by material ejected in its nineteenth century eruption. The physical cause of this event is still entirely unknown.

6.5 The final countdown...

The previous sections showed that the post-main sequence evolution of low-mass and high-mass stars is similar: when fuel runs out, the core contracts and heats up, and this makes it possible for new nuclear reactions to burn new sources of fuel. For stars similar to the Sun, this ends with the slow ejection of their outer layers to leave behind a hot but inert core made of carbon and oxygen.

However, in contrast to this stately progress, the death of massive stars is much more rapid and violent. Section 6.4 showed you the nuclear processes after the fusion of helium that supply the energy to support a star against gravity. In other words, this supply of energy prevents such a star collapsing inwards under its own weight.

But these reactions aren't all equally efficient. Each successive chain of reactions, involving progressively heavier elements (carbon-, neon-, oxygen- and silicon-burning) is *less efficient* at releasing energy than the preceding one. This means that each reaction chain must consume more fuel per second than the previous one to produce the same amount of energy to support the star against gravity. This means that each phase of burning will be of shorter duration than the one before it. For example, in a star 25 times as massive as the Sun, the hydrogen-burning phase lasts 7×10^6 years and helium burning lasts 5×10^5 years. Subsequently, the carbon-burning phase only lasts about 600 years, neon burning lasts about 1 year and oxygen burning only 6 months. Finally, silicon burning takes only about 1 day until the star runs out of this fuel source.

What happens next? As before, in the absence of an energy source, the core collapses under gravity, with a resultant rise in temperature to the point at which iron nuclei can undergo nuclear fusion. But this is where things change. The fusion of iron to form heavier elements is different from that of lighter elements in one vital respect. rather than releasing energy it *absorbs* it. The effects of this are dramatic. Nuclear fusion can no longer supply the energy to support the star against gravity. Therefore, the core continues to collapse in upon itself. Soon it reaches the densities at which electron degeneracy pressure stops the contraction in lower-mass stars. But in more massive stars the force of gravity quickly overwhelms the resistance of electron degeneracy pressure. It's predicted that electron degeneracy pressure can only fully support cores of stars with masses of about 1.4 times that of the Sun, but some stars have much more massive iron cores than this. The collapsing core continues to rise in temperature until it reaches the point at which the iron nuclei begin to break apart into their constituent protons and neutrons. Again this process requires an input of energy, so it deprives the star of energy needed to support its own weight and thus helps accelerate the collapse. As the core density increases so does the electron degeneracy pressure that resists the contraction until the negatively charged electrons and positively charged protons combine to form neutral neutrons. This removes the electron degeneracy pressure so the core collapse gains speed.

The core collapse suddenly stops at the staggeringly high temperature of 10^{12} K and density of approximately 3×10^{17} kg m^{-3} when a new form of degeneracy pressure, **neutron degeneracy pressure**, comes into play. This works on exactly the same quantum principles as electron degeneracy pressure, except neutrons are the particles involved instead of electrons. The pressure provided by the neutrons is enough to halt the collapse of the core, but the remaining outer layers of the star are still falling inwards at speeds of up to 70 000 km s^{-1}, or over 20% of the speed of light! What happens next is uncertain. Even the most advanced computer simulations cannot accurately reproduce the process, but the energy of the infalling material is redirected outwards as the outer layers hit the core and rebound, expelling them in a titanic explosion called a **supernova** (e.g. Figure 6.6).

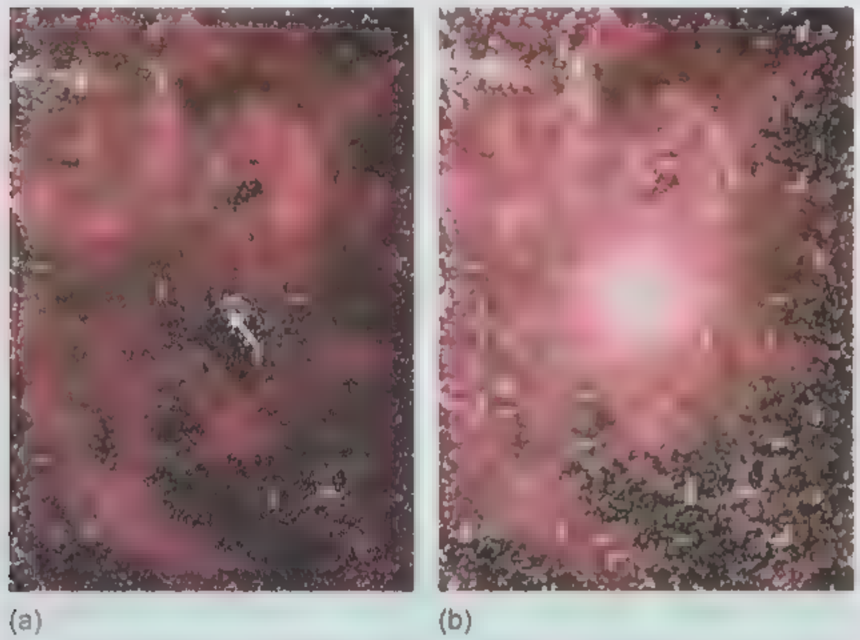


Figure 6.6 Before (a) and after (b) pictures of the supernova SN 1987A which occurred in the nearby Large Magellanic Cloud. The progenitor of the supernova is thought to be a blue supergiant star roughly twenty times as massive as the Sun. Images courtesy of the Anglo–Australian Observatory.

Whatever the cause of the explosion, the high temperatures present and abundant supply of neutrons allows the star to undergo one final round of nucleosynthesis. In only a few seconds the neutrons are rapidly absorbed into nuclei in the outer layers of the star. This produces a wealth of new heavy elements including gold and plutonium. These are called **r-process reactions**, with the 'r' standing for rapid.

The star spent millions of years evolving to the point at which it has a massive iron core, but its final death throes are over in a matter of seconds. So much energy is released in this final explosion that the star brightens by a factor of about a hundred million. For a while it might outshine the *entire galaxy* in which it is located. If one were located in our own galaxy it would be visible during the day for a few weeks!

The star has one final gift for the cosmos: the explosion returns all the elements that were formed within it by nucleosynthesis back into the galaxy, into the interstellar gas and dust located between stars (see Figure 6.7 overleaf). The next generation of stars will then form from this material. Supernovae are the most important sources of elements heavier than neon in the Universe. The iron found in our blood and in the molten core of the Earth is the direct result of the explosive death of a massive star.

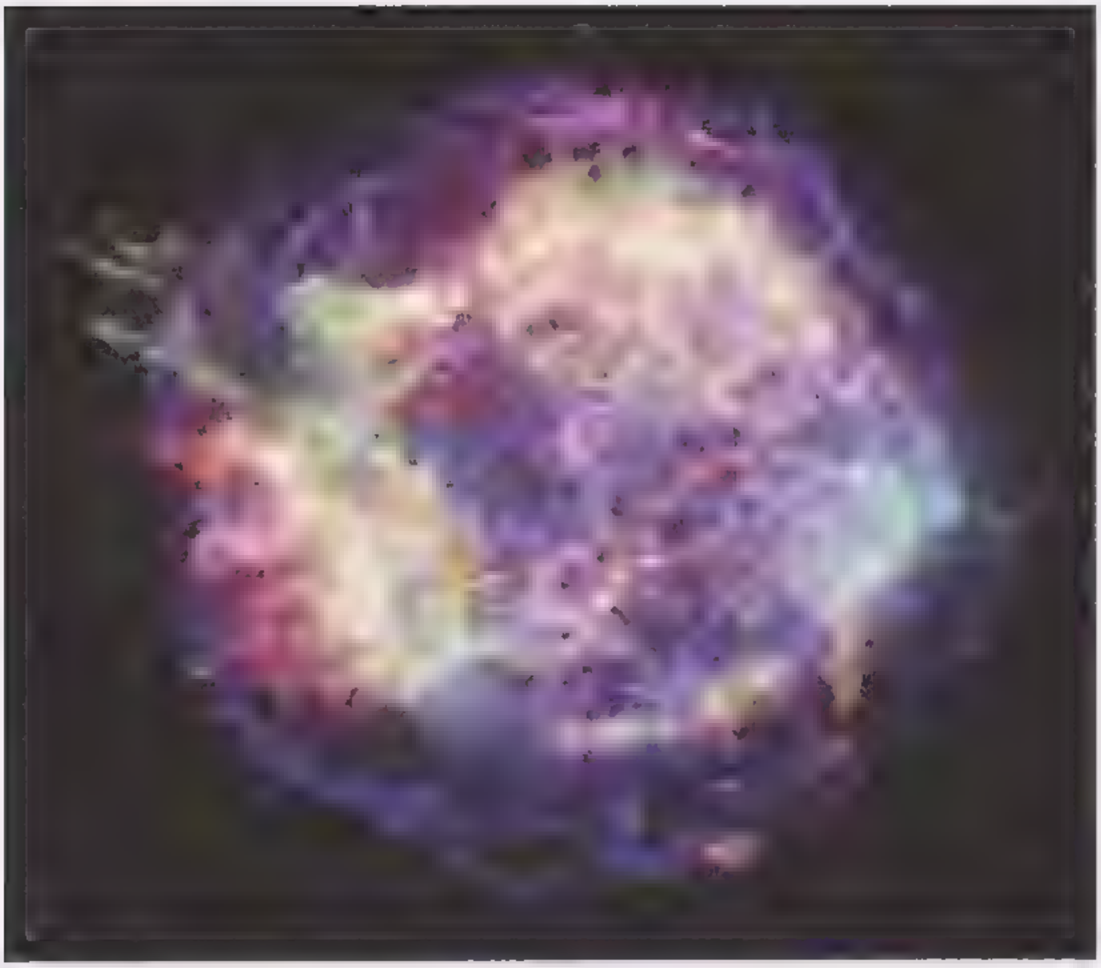


Figure 6.7 X-ray image of the Cassiopeia A supernova remnant, highlighting the hot chemically enriched material returned to the Galaxy.

6.6 The final remnants of stars

The previous section looked at the final life stages of stars with masses similar to the Sun, and stars much larger than the Sun. These final life stages differ because of the different physical properties of their cores and the nuclear reactions that occur there. However, they also share similarities: in both cases they eject large quantities of material enriched with the products of stellar nucleosynthesis back into the rest of the galaxy. This process doesn't lead to the complete destruction of the star, because a dead remnant of the progenitor star is left behind. But what are the properties of such stellar corpses?

6.6.1 White dwarfs

White dwarfs are the remains of stars such as our Sun. After the outer layers of the star are ejected, the hot stellar core is exposed. This core is composed of the carbon and oxygen 'ash' of previous episodes of nuclear burning. The lack of a current energy source means that these stars are now entirely supported against gravity by electron degeneracy pressure. *Newborn* white dwarfs are still very hot, because nuclear reactions have only recently ceased. The surface temperatures of the most extreme examples known reach

150 000 K. Most are considerably cooler though, with temperatures around 25 000 K. The coolest known are as low as 4000 to 5000 K.

Given this, you may be surprised to find that white dwarfs are rather faint (e.g. Figure 6.8 and see also Figure 6.2); their location on the HR diagram shows that they are typically less luminous than the Sun. The combination of high temperatures and low luminosities must mean that they are physically rather small, and it turns out they're typically about the same size as the Earth! Despite this, they still contain a large amount of the matter that made up the original star. Measurements of the masses of white dwarfs indicate that they have masses ranging from 17% to 133% of that of the Sun. They must therefore be extremely dense objects, with average densities thousands of times higher than the Sun, or about a thousand, million kilograms per cubic metre!

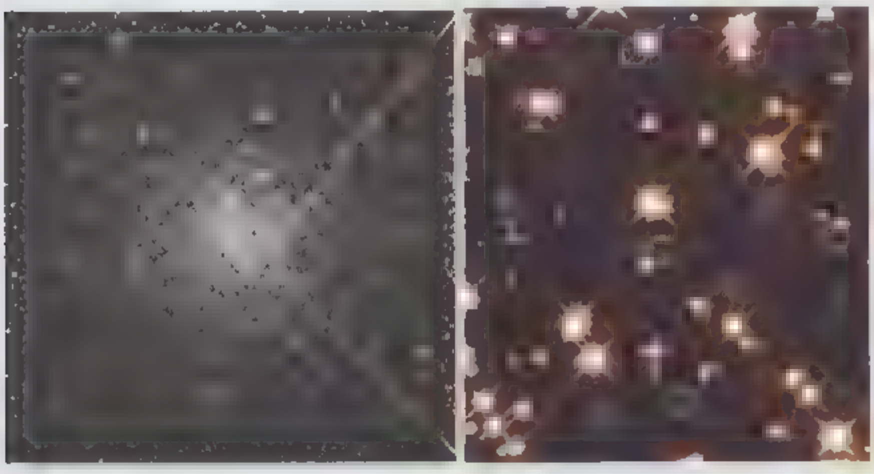


Figure 6.8 The results of a search for white dwarf stars in the globular cluster M4, located about 2 kpc from the Sun and containing more than 10^5 stars. The left-hand panel shows the entire cluster as viewed from ground-based telescopes. The area searched by the Hubble Space Telescope is marked. The right-hand image shows the eight isolated white dwarf stars (circled) that were found. The other objects present are main-sequence stars and red giants.

What are the ultimate fates of these objects? With no more sources of internal energy available to them they will gradually radiate their residual heat away into space, cooling and becoming progressively dimmer as they do so. White dwarfs have much smaller surface areas than other stars, so it takes a long time for all the energy of such a star to be radiated. This would take about as long as the current age of the Universe (about 10^{10} years). The unavoidable fate of stars like the Sun is that they will end their days as cold, dark spheres rich in carbon and oxygen, but lost to the cosmic cycle.

6.6.2 Neutron stars

Section 6.5 showed you that above a mass of about 1.4 solar masses (meaning 1.4 times the mass of our Sun), electron degeneracy pressure is overwhelmed by the force of gravity and the star collapses in upon itself again. This boundary is called the **Chandrasekhar limit**, in honour of its prediction by the Indian astronomer Subrahmanyan Chandrasekhar in 1930. His prediction has compelling support from a simple observational fact: in almost a century, we have yet to find a white dwarf with a mass above this value.

During this collapse, the electrons and protons combine to form neutrons. The collapse is only halted once the neutrons themselves are so close together that neutron degeneracy pressure halts the contraction. By this stage the star has shrunk to a radius of only about 10 km and is made predominantly of neutrons. For this reason it's called a **neutron star**. Neutron stars have densities of 4–7 × 10^{17} kg m^{-3}. These mind-bogglingly high densities are similar to those found at the centre of atomic nuclei and are a factor of a hundred million times greater than even those found in white dwarf stars. A single thimble-full of neutron star matter would have a mass of an astonishing seven hundred million tonnes. These huge densities mean that neutron stars also have powerful gravitational fields. If an object were dropped from 1 metre above the surface of a neutron star, it would hit the surface a microsecond later travelling at 2000 km s^{-1}!

For any astronomical object, it's possible to define an **escape velocity**. This is the speed at which something must be moving away from it to completely escape the object's gravitational field (Section 4.7). For the Earth this is 11.2 km s^{-1}; this is the speed you would have to kick a ball for it to escape Earth's gravity (neglecting air resistance). For neutron stars, the escape velocity is 30% of the speed of light!

However the surprises don't stop there. Directly after the formation of the neutron star, its surface temperature is more than 10^8 K. Like white dwarfs they don't have any internal source of energy and so they are relentlessly cooling. Within a few thousand years they will have reached 'only' 10^6 K and after around a million years they'll be at about 10^5 K. These temperatures imply that most of the energy will be radiated away as X-rays.

Another extreme property of neutron stars is the rate at which they rotate, typically completing several full rotations every second. Some rotate over 700 times per second! Why? The reason can be appreciated by watching a skater spinning. As they pull in their arms, the rate at which they spin increases. Exactly the same phenomenon takes place during the collapse which leads to the formation of a neutron star: the comparatively slow rotation of their progenitor star rapidly increases as the star contracts by a factor of about 10^5.

At both X-ray or radio wavelengths the emission from neutron stars arises from 'hot spots' on its surface. This makes it possible to measure the rotation rates of neutrons stars directly. Every time an X-ray-bright or radio-bright region of a neutron star rotates into our line of sight, astronomers see a bright flash at X-ray or radio wavelengths. The result is a regular series of pulses that make it possible to measure their rotation rate. They also give neutron stars their alternative name: **pulsars** (Figure 6.9a).

Activity 6.1 The properties of a pulsar

The estimated time for this activity is 10 minutes.

All you need to see for yourself how a slowly rotating massive star can form a rapidly rotating neutron star, or pulsar, is a rotating stool or chair (e.g. an office chair). Clear some space around it and then get someone to spin you round with your arms and legs outstretched. Now if you bring your arms and legs into your body you will find that you spin more rapidly – exactly the same physical effect that takes place as a pulsar is born!

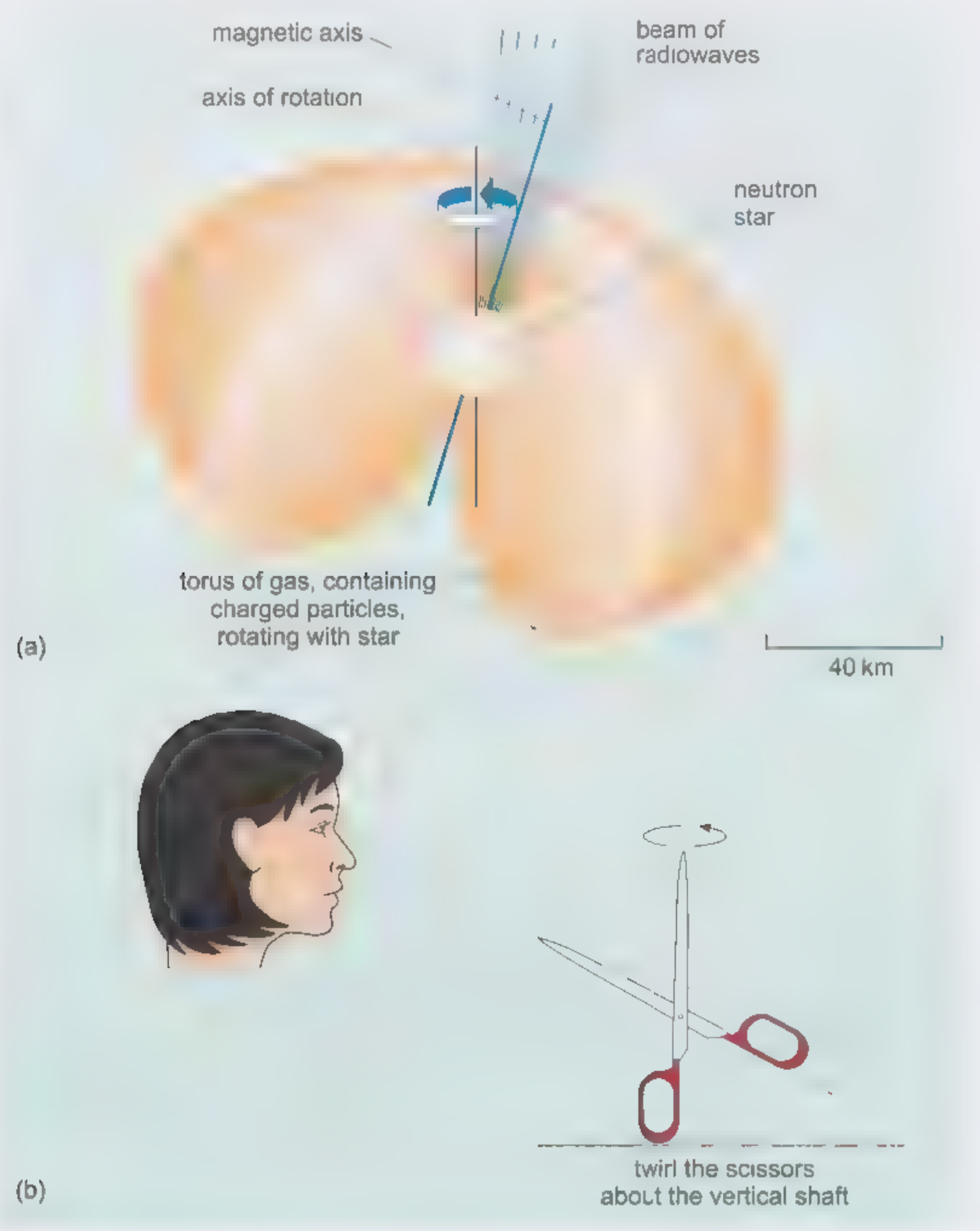

magnetic axis
beam of radiowaves
axis of rotation
neutron star
torus of gas, containing charged particles, rotating with star
(a)
40 km
(b)
twirl the scissors about the vertical shaft

Figure 6.9 (a) An artist's impression of a pulsar. In reality, the torus (doughnut-shaped ring shown here in cutaway view) of gas would extend further than shown here. (b) Modelling the way we 'see' radio pulses from a pulsar.

To understand the 'pulsing' of a pulsar, all you need is a pair of scissors (preferably with straight blades and rounded ends (as shown in Figure 6.9b) and some Blu-Tack® or plasticine Use the Blu-Tack® to keep the blades open and twirl the scissors about one shaft which is held upright (being careful not to stab yourself) Experiment until you find an arrangement where the slanted blade points directly at you once in each revolution. This slanted blade represents the radio beam which you can ‘see’ only when it is directed towards you, so you observe a series of short flashes as the pulsar rotates.

Above the Chandrasekhar limit, the force of gravity overwhelms electron degeneracy pressure in white dwarfs. Is there similarly an upper limit to the mass of neutron stars above which neutron degeneracy pressure is unable to support them? It’s believed that the answer is yes, but current theories do not supply an unambiguous value. The reason for this is that no-one yet understands how material at such extreme high densities behaves. This is mainly because there are no current experiments that can produce such extreme conditions. It’s strongly suspected that neutron stars can’t be more massive than 5M and that the real limit is probably somewhere between 2 and 3 times the mass of the Sun.

Some scientists instead turn this problem round, and use the masses of neutron stars as a way of improving their understanding of dense matter! Neutron star masses can be measured in binary systems, in exactly the same way that the masses of stars powered by nuclear fusion and white dwarfs are measured: the faster they orbit around one another, the more massive they are. At the time of writing (early 2012), the most massive neutron star discovered so far (with a reliable mass measurement) has a mass of 1.94 ± 0.04 times the mass of our Sun. This means that the star’s mass could be as little as $1.90\,\mathrm{M}_\odot$ or as large as $1.98\,\mathrm{M}_\odot$ or any value in between.

6.6.3 Black holes

Suppose the force of gravity causes gas to fall onto the neutron star, pushing it over the limit. Once the force of gravity overwhelms the ability of neutron degeneracy pressure to support a neutron star, is there another similar force that will act to support it and prevent further collapse? To the best of current knowledge, the answer to this question is no. At this point there is no further physical force that can prevent the further, final collapse of the neutron star under its own mass. There seems to be nothing to stop it shrinking to zero size and infinite density. What happens then is somewhat uncertain because all current theories break down at the point of the most extreme densities. What’s needed to solve this problem is a theory that combines Einstein’s theory of gravity, general relativity, and the theory that governs the behaviour of the Universe on very small scales, quantum mechanics. So far, no-one has been able to make a working version of such a theory. But there is one thing all the current theories are clear on: as the collapse continues the core will become so dense that the escape velocity from the surface will exceed the speed of light. The speed of light appears to be the ultimate speed limit for the Universe[1], so this means that nothing, not even light, can escape the gravitational pull of the

[1] At the time of writing (early 2012) there have been some claims that neutrino particles can sometimes travel faster than light. These claims are being intensively debated by scientists and there are indications that a faulty cable led to spurious results. *If* these results are nevertheless right, many things that were believed to be correct about the Universe are in fact wrong. This would be intriguing and wonderful. However the level of scepticism is such that the physicist Professor Jim Al-Khalili announced he would eat his shorts on live television if the results turn out to be right.

core. At this point a **black hole** will have formed. The distance from the centre at which the escape velocity is the speed of light is called the **event horizon** and its size depends on the mass of the collapsing core. Once an object crosses the event horizon it is forever trapped within the black hole. The event horizon doesn't describe a solid surface, any more than the horizon you'd see from a ship is a solid line. An unlucky astronaut could pass over the event horizon of a sufficiently massive black hole without noticing any ill effects and yet be trapped inside forever by the gravitational field of the black hole.

How can anyone find black holes? No light can escape from black holes, so it might seem impossible in principle to observe them, but they can have a detectable influence on nearby objects. This can happen if a normal star orbits around a black hole. In such a binary star system the intense gravitational field of the black hole tears material from the surface of the star, which then falls into the black hole. As the material falls in, its gravitational potential energy is converted into heat energy, so it heats up to very high temperatures. This process occurs *outside* the event horizon, so this energy can be radiated away as X-rays, even though the material subsequently falls inside the black hole.

Therefore, one way of finding binary star systems containing black holes is to identify very bright X-ray sources co-incident with normal stars. Astronomers then measure the motion of the normal star as it is thrown around its orbit by the invisible black hole, and from that determine its mass. Neutron stars can't be more massive than $5\,M_{\odot}$, so if the invisible companion is more massive than that, then it's a black hole.

The best candidate for a black hole has been found at the very centre of our own galaxy. Astronomers have been monitoring the motion of stars in the centre for many years with some of the largest telescopes. They are orbiting a central object, but no luminous object is to be found there. Whatever it is, it is certainly very small and very massive, because the stars move very quickly around it. It is so dense it seems impossible for it *not* to be a black hole. Its measured mass at the time of writing is 4.31 ± 0.38 million solar masses!

Finally, how are black holes produced in the first place? One way is if enough material falls onto a neutron star (or is *accreted*) from a companion star, it could exceed the maximum mass for a neutron star, so it would then collapse to form a black hole.

It's also possible for very massive *single* stars to form black holes. Two possible avenues for this have been suggested. If the star is massive enough, the supernova explosion that forms the neutron star might not be able to release enough energy to eject all the outer layers of the star. That material will then fall back onto the newly-born neutron star, causing it to collapse to form a black hole. Unfortunately, supernova explosions are not yet well understood, so it's not possible to give a hard and fast value for the mass of a star above which this would occur, but the best estimate at the moment is that stars above 25 times the mass of the Sun will form black holes via this route.

Alternatively, could the mass of the pre-supernova iron core be so high that it collapses directly to form a black hole, without forming a neutron star or undergoing a supernova? This might work for extremely massive stars, perhaps greater than a hundred times the mass of the Sun. To date no firm observational evidence supports such a prediction, but perhaps over the next couple of decades some new all-sky surveys by robotic telescopes might identify such 'vanishing' massive stars.

6.7 The cosmic cycle and the next generation of stars

This chapter and the preceding one studied the life cycle of stars, emphasising the role nuclear fusion plays in both powering stars and creating new heavy elements via nucleosynthesis. As these stars die, these newly forged elements are returned to the interstellar environment of the Galaxy, known as the **interstellar medium**. This can happen either gently through stellar winds or explosively in supernovae. The origin and composition of the interstellar medium is illustrated graphically in Figure 6.10, which shows the relative abundance of each element and its physical origin. Hydrogen (and its isotope deuterium), helium and lithium formed in primordial nucleosynthesis in the early Universe, but the remaining elements are *all* the product of nuclear fusion reactions within stars and supernovae. But how do these elements become recycled and incorporated into the next generation of stars? This last section of the chapter will answer this by investigating how stars form, closing the loop in the cosmic cycle.

Outer space is *not* a pure vacuum, contrary to popular perception; the matter is just much, much more sparse than any vacuum chamber can achieve on Earth. Some of the contents of the interstellar medium are visible as the dark lanes in the images of the Antennae galaxies (Figure 6.1). The dark lanes are regions of obscuration from **dust** in the interstellar medium. This dust is *very* different from household dust. It's composed of carbon-rich molecules and tiny mineral-rich grains about 1 micron in size, containing elements such as oxygen, iron and silicon.

The interstellar medium has a surprisingly wide range of temperatures and densities: for example, gas near a cluster of very massive stars will be at a higher temperature than more isolated material. Nevertheless, regions of the interstellar medium containing both gas and dust will be at comparatively 'low' temperatures. This is known because the dust will melt and vaporise if temperatures exceed about 2000 K. The temperatures of the dense clouds of dust and gas where stars form are *much* cooler, at only a few tens of kelvin. These regions are known as **Giant Molecular Cloud complexes**, or **GMCs**. They are large and massive, extending across about 100 pc and containing more than a million solar masses of gas and dust. While their densities are very high compared with the rest of interstellar space they are nonetheless very diffuse, with densities 10^{15} (a thousand million million) times lower than the atmosphere of the Earth at sea-level. However they contain so much dust that they can completely absorb all the optical light from stars behind them (Figure 6.11). The presence of young stars near GMCs strongly suggests that they're intimately involved in star formation. But how are these large, massive, diffuse clouds converted into stars?

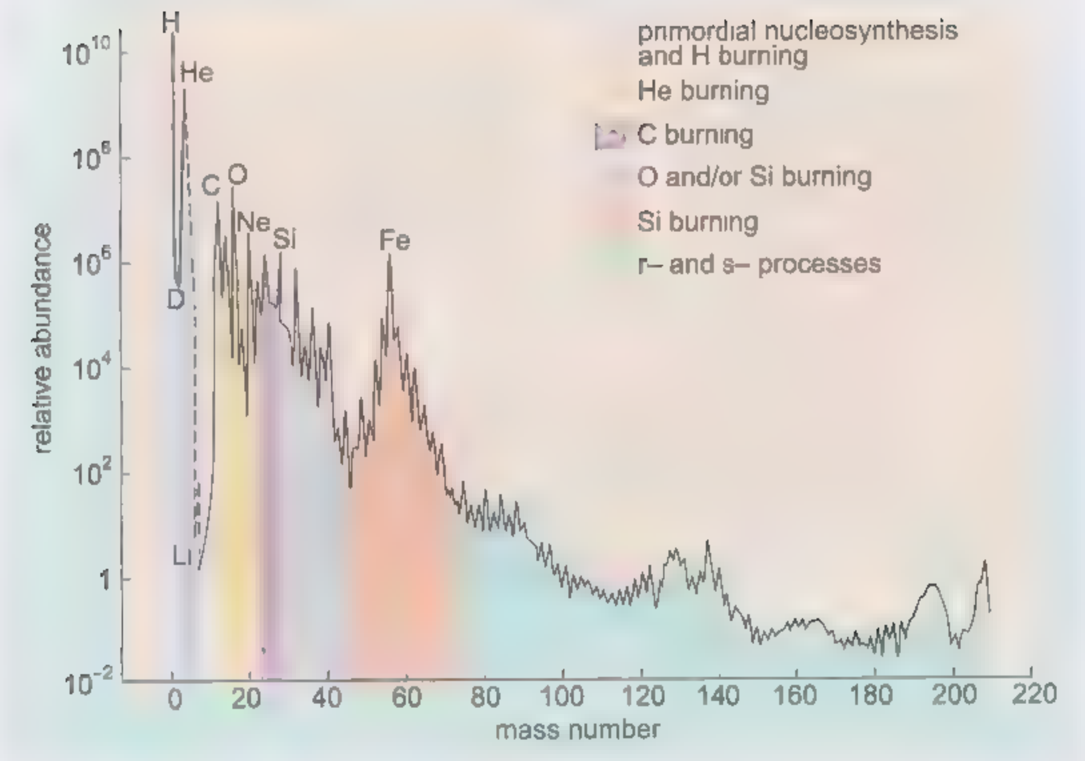


Figure 6.10 The distribution of elements by relative abundance in the local interstellar medium. The stages of nuclear burning that give rise to elements at different mass numbers are indicated.

Three key ingredients seem to be needed for star formation to occur. Firstly, the GMC is highly structured, with lots of clumps and filaments having much higher densities than other regions of the cloud. These are the locations where star formation will occur. The second ingredient is gravity, which will cause the clumps to collapse in on themselves. However, these clumps are initially supported by their thermal energy (in exactly the same way that stars are supported against gravity), so the final ingredient is something to disturb this equilibrium and trigger a collapse. The trigger is an external event that compresses the clumps so that gravity can take over and cause them to begin their contraction. This can be something like a nearby supernova, where the shockwave from the explosion hits the GMC, or perhaps the collision of two GMCs.

Figure 6.11 Picture of a dense molecular cloud called a Bok globule. The density of dust and gas is sufficiently high that it absorbs light from stars located behind it.

Once a clump within a GMC is sufficiently dense it will start to contract under its own gravity. As it contracts, the conversion of gravitational potential energy to heat energy causes the temperature of the molecular clump to rise, again in exactly the same way as the core temperature of a star rises as it contracts. This is depicted in Figure 6.12, which shows a GMC about 2 kpc from the Earth in the constellation Carina. The right-hand panel shows this region at optical wavelengths where the cold dust absorbs background light, causing the dark V-shaped absorption feature across the bright nebula. At infrared wavelengths, the thermal emission from this cold dust can be detected. The infrared image in Figure 6.12 shows the true extent of the GMC. There is bright infrared emission from the cold material that corresponds to the dark V-shaped lane in the optical. (The bright material visible in the optical image is gas being heated by newly born stars outside the GMC.)

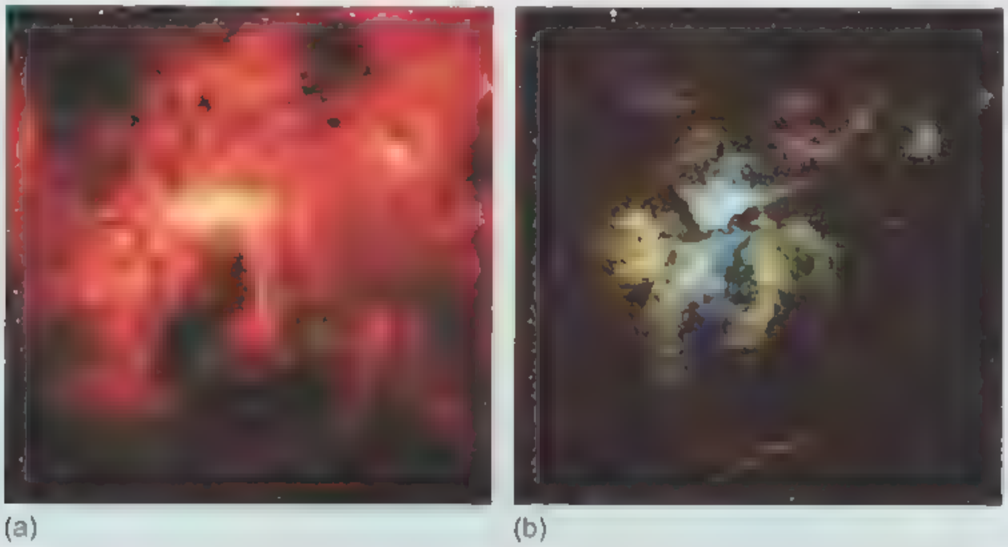


Figure 6.12 Comparison of the Giant Molecular Cloud associated with vigorous recent star formation in the constellation Carina at (a) infrared (b) and optical wavelengths. Note the dark V-shaped lane in the optical due to cold dusty material is observed to strongly emit infrared (IR) radiation due to heating from the massive protostars embedded within.

The contraction of the clump is expected to be remarkably rapid. Computer simulations of this process suggest that after only a few thousand years the surface temperature of the contracting clump, or **protostar**, will have risen to 2000 to 3000 K! This is beautifully illustrated by the observations of a small region of the Carina GMC in Figure 6.13, where a hot young protostar is clearly picked out by the light it is radiating at infrared wavelengths. In addition to heating up as they collapse, the cloud clumps also begin to spin more rapidly. This happens for exactly the same physical reason that you saw in the formation of neutron stars. As they spin more rapidly they change their shape, becoming flatter and more disc-like Material falling onto the central protostar now passes through a flattened region surrounding it called an **accretion disc**. At the same time as material is falling onto the equatorial regions of the protostar, powerful jets of material are ejected from the poles of the star. This simultaneous infall and outflow geometry is shown schematically in Figure 6.14 and observations of these phenomena are shown in Figure 6.15. Astronomers are currently unsure of the detailed processes that launch these jets, but their effect is to remove both material and energy from the protostar.

Eventually, the core of the protostar will become so hot that nuclear fusion reactions can begin and a new star is born. This whole process is surprisingly rapid and is thought to be complete within 10^8 years for the lowest mass stars. More massive protostars have a greater gravitational pull, so material is pulled onto them at a much higher rate, which makes more massive stars form more quickly. Stars of about $15 M_{\odot}$ form in only 10^5 years. In any circumstances, the pre-main sequence phase of a star is expected to be much shorter than its lifetime on the main sequence.

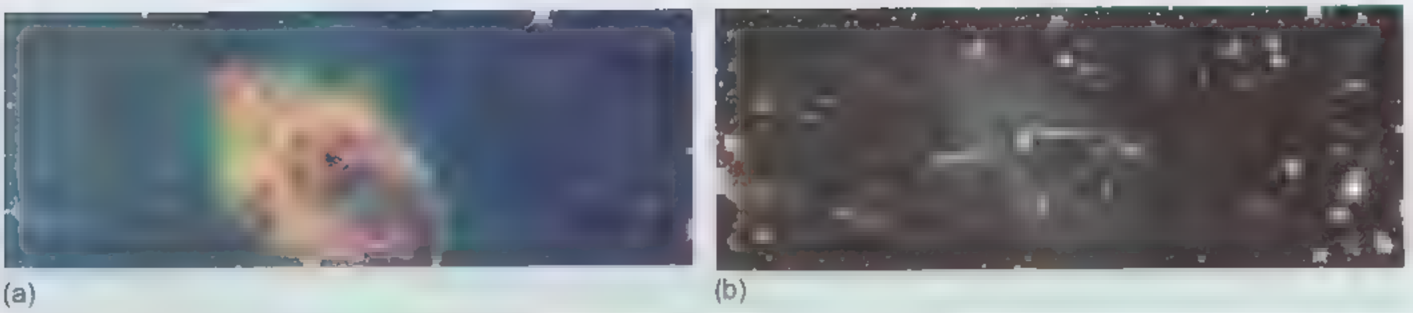


Figure 6.13 A close-up of a region of the Carina GMC. In the optical image (a) we can see the opaque molecular cloud while in the near-IR image (b) we can clearly see the protostar which is in the process of formation. Note also the bright linear jets of material associated with the protostar.

Astronomers are increasingly confident of this picture of star formation for stars of mass comparable to the Sun. By contrast the formation of very massive stars is still not fully understood. Stars like these are found in clusters such as the Arches (Figure 5.5) in the centre of our galaxy. The formation is quick and the stars are rare, both of which make the observations difficult. Also, the GMCs they form in are expected to be so dense that little light can escape from them during their formation. Perhaps they form in essentially the same way as lower-mass stars, but there's an alternative theory of 'stellar cannibalism'. In this theory, smaller protostars collide and merge to rapidly build up a massive protostar. This merger process would become more and more efficient as the central mass and gravitational pull increase with every merger.

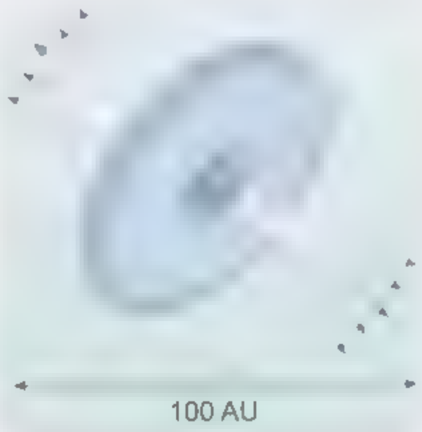


Figure 6.14 Schematic representation of a protostar, surrounded by an accretion disc which supplies *infalling* gas to its equatorial regions. Also shown are *outflowing* jets from the polar regions, which remove both energy and a proportion of the inflowing material.

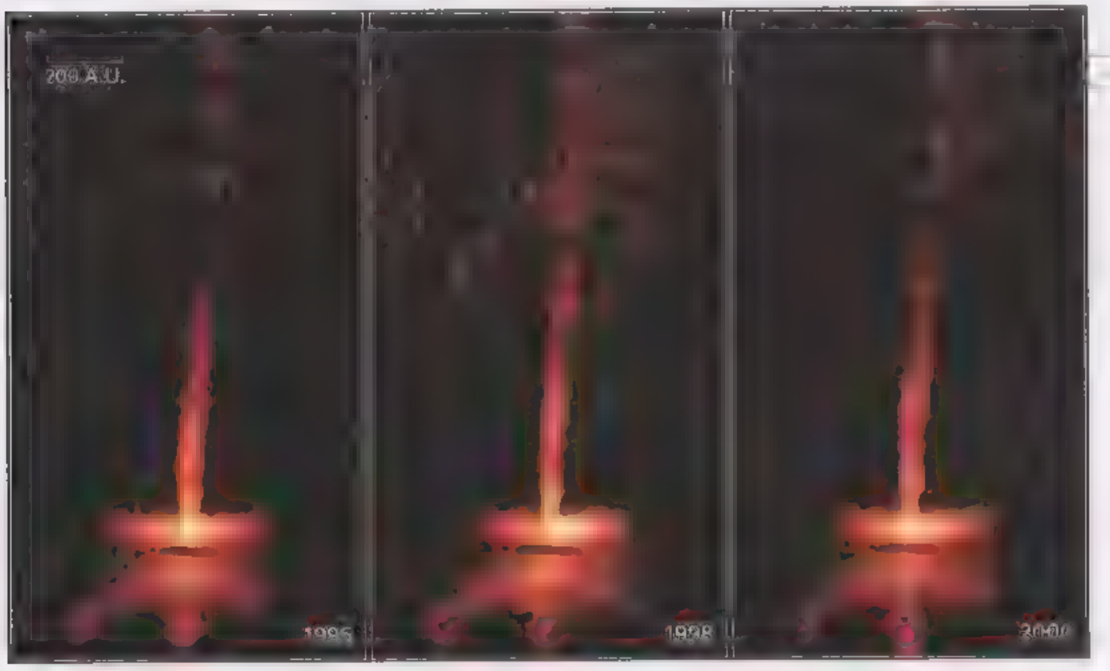


Figure 6.15 Sequential images of the protostellar object HH30 taken in 1995, 1998 and 2000 illustrating the features shown in the schematic in Figure 6.14. The densest regions of the horizontal, edge-on disc hide the protostar itself, the light from which illuminates the outer surfaces of the disc. The motion of blobs of gas within the jets can clearly be seen.

These last two chapters have shown you how stars of all masses are born, live and finally die. All these phases are intimately connected: nuclear fusion creates new elements in stars, which are subsequently returned to the interstellar medium as the stars die, ready to be incorporated into the next generation of star formation. This continuous cosmic cycle is shown schematically in Figure 6.16 and is dramatically in evidence in the beautiful Hubble Space Telescope image of the young stellar cluster NGC 3603 (Figure 6.17). However, this is not *quite* the end of the story, because the accretion discs around protostars hide one final secret. Accretion discs are the sites where new planetary systems are born, and this is the subject of the next chapter.

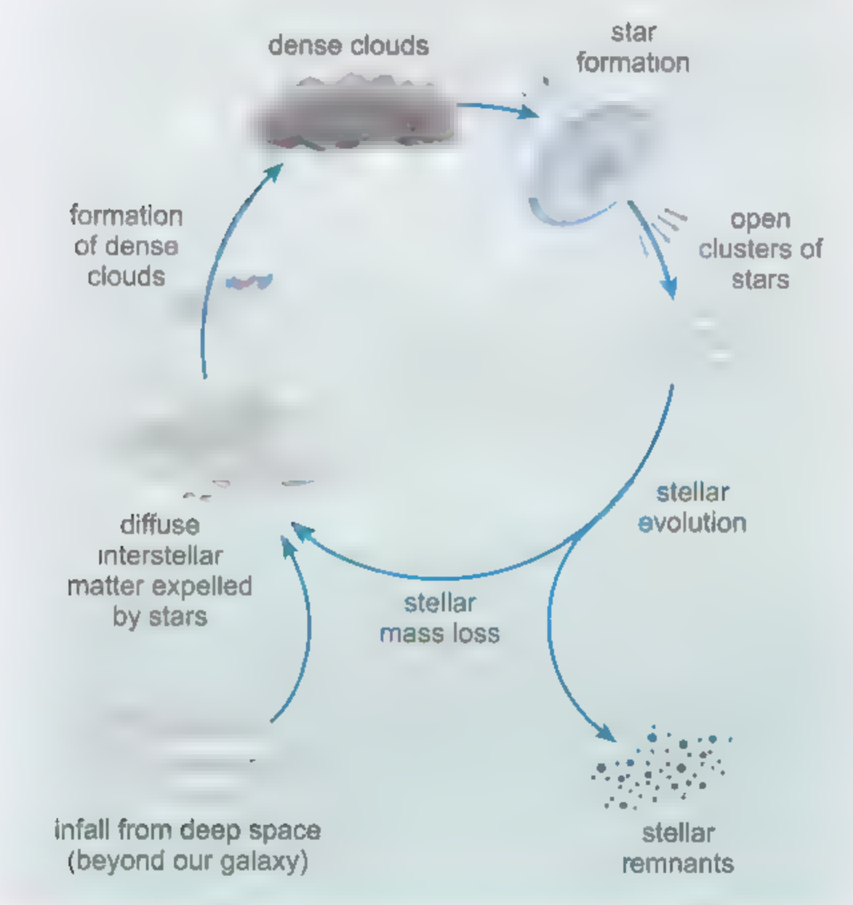

dense clouds
star formation
open clusters of stars
stellar evolution
stellar remnants
stellar mass loss
infall from deep space (beyond our galaxy)
diffuse interstellar matter expelled by stars
formation of dense clouds

Figure 6.16 The cosmic cycle.

End-of-chapter questions

Question 6.1 From the following terms select six and place them in the correct chronological order to describe the life cycle of a star like the Sun:

Supernova, dense cloud, planetary nebula, neutron star, red giant, supergiant, white dwarf, protostar, main-sequence star

Question 6.2 In this chapter you met various types of nebulae. Using the Web find images of the Eagle nebula, the Ant nebula and the Crab nebula and for each one write a sentence or two describing how it illustrates a stage of the life history of a star.

Question 6.3 When astronomers look at our galaxy they see many more main-sequence stars than young protostars. Why is this?

Question 6.4 If instead of releasing energy, the fusion of hydrogen to form helium absorbed energy, what might happen to stars?

Now go to the module website and do the remaining activities associated with Chapter 6.

Activity 6 2 - 20mins
Activity 6.3 - 40mins
Activity 6.4 - 40mins
Activity 6 5 - 40mins
Activity 6 6 - 60mins

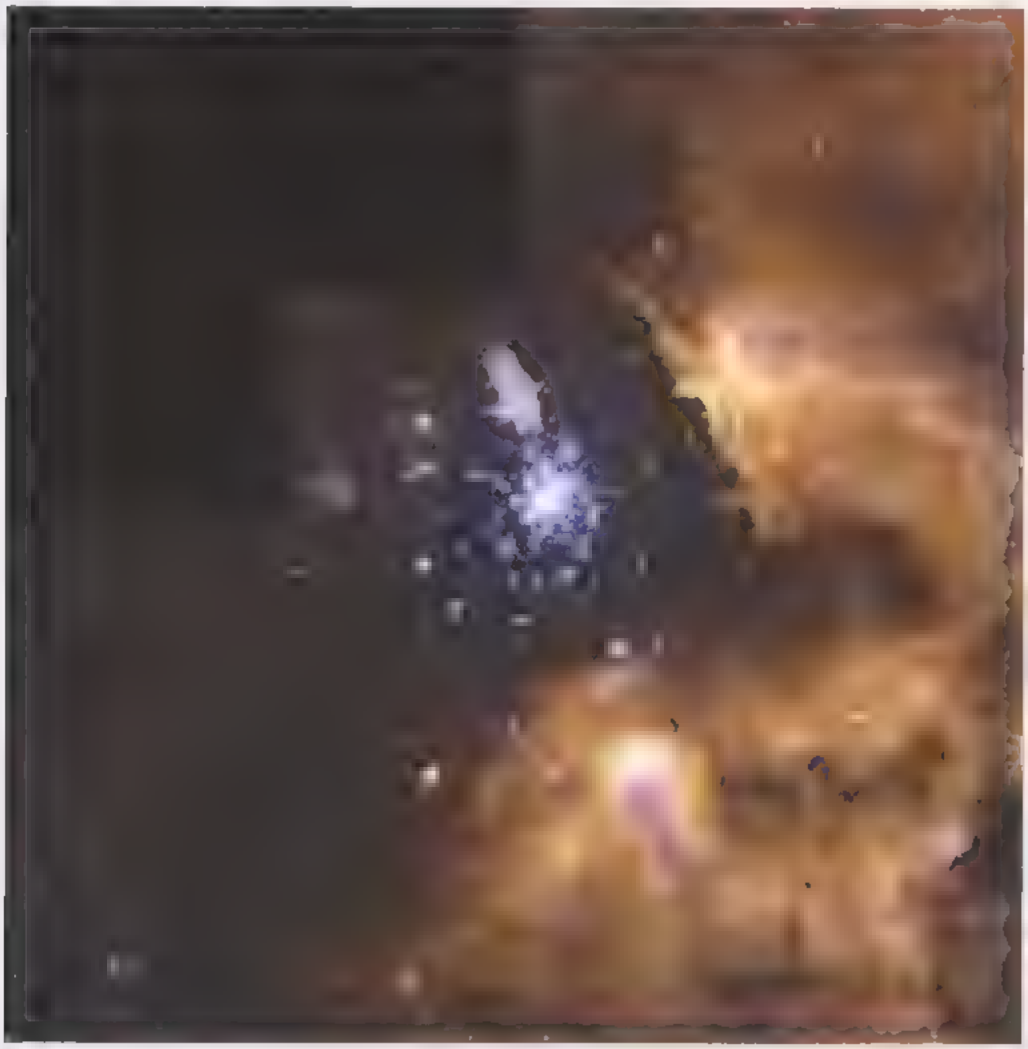


Figure 6.17 Hubble Space Telescope image of the young massive cluster NGC 3603 and associated GMC, illustrating the cosmic cycle The most massive stars in this region are already beginning to die. Note the material ejected from the blue supergiant in the upper left, returning the products of nucleosynthesis to the cosmos. (Note that the black patch in the very top left corner of the image indicates that data were not available from this area to add to the image.) The light and winds from the young stars in the cluster are compressing the surface of the GMC from which it was born, leading to a further burst of star formation at the bright heads of the pillars within it.

Chapter 7 Planetary systems

7.1 Introduction

The Solar System, including the Sun and its retinue of planets and smaller bodies was formed around 4.6 billion years ago from a collapsing cloud of interstellar gas and dust. As you have learnt in the previous chapter, this cloud contained atoms of heavier elements produced during the lives and deaths of stars. It is these atoms that form the majority of the mass of the terrestrial planets and play a key role in the development of life (see Table 1.2).

This chapter reviews the properties of the Solar System (Section 7.2), the different objects it contains (Sections 7.3 and 7.4) and their role in its evolution (Section 7.5). While it is only one example of a planetary system, the Solar System is the only available source of detailed information on the properties of planets and how they form and evolve Before 1992, no planets beyond the Solar System (known as **extrasolar planets** or **exoplanets**) were known. Although their existence, and the possibility of life on other worlds, had been speculated on for centuries, the techniques to detect exoplanets have only been available for a few decades. Many hundreds of planets are now known, but most of the techniques presently available for their detection (Section 7 6) are most likely to find large planets close to their host star where conditions are unsuitable for the kind of life found on Earth. Section 7.7 investigates the properties of known exoplanets and the possibility of finding Earth-like planets.

7.2 Survey of the Solar System

7.2.1 Orbits in the Solar System

Broadly speaking, the Solar System consists of objects orbiting other objects in more-or-less circular paths. The planets orbit the Sun and they, in turn, have satellites and rings in orbit around them.

■ Why can the orbits of Solar System objects be described as 'well-ordered'?

The orbits of the planets lie in (almost) the same plane, and they all move around the Sun in the same direction.

As you will see later in this chapter, this well-ordered motion helps astronomers to explain the origin of the orbital motion of planets and their satellites.

The planets and other bodies in the Solar System do not, however, orbit in exact circles. The German astronomer Johannes Kepler (1571–1630) formulated three laws of planetary motion to describe the motions of planets These laws are fundamental to our understanding of planetary motion and apply to all bodies orbiting a central body under the influence of its gravity Kepler's first two laws relate to the shapes of orbits and the speed at which a body travels in its orbit. Although most of the planets orbit in almost circular orbits, comets and some asteroids can have highly elongated orbits. The degree of elongation of an orbit can be described by its **eccentricity**, *e*. Eccentricity can have values between 0 for a circular orbit and close to 1 for a highly elongated orbit (Figure 7.1). Objects in circular orbits travel at a constant speed around the central body, but objects in elliptical orbits move faster when they are closer to the central body and more slowly when they are far away. A comet with a highly elongated orbit will spend most of its time far away from the Sun.

Kepler's third law relates the average distance and orbital period of a planet and allows the mass of a star (or, more correctly, the combined mass of the star and the planet) to be calculated. This is because the average distance of the planet from the star a is related to its orbital period P by the equation:

$$P^2 = \frac{ka^3}{M+m} \tag{7.1}$$

where M is the mass of the star, m is the mass of the planet and k is a constant. This is a mathematical representation of Kepler's third law. It applies equally to circular or elliptical orbits.

- If we take the case of the Earth orbiting the Sun, and measure the orbital period in years, the average distance of the Earth from the Sun in astronomical units, and the mass of the Sun in solar mass units ($\mathrm{M}_\odot$) what is the value of k?

If P = 1 yr then

$$P^2 = 1\ \mathrm{yr} \times 1\ \mathrm{yr} = 1\ \mathrm{yr}^2.$$

If a = 1 AU then

$$a^3 = 1\ \mathrm{AU} \times 1\ \mathrm{AU} \times 1\ \mathrm{AU} = 1\ \mathrm{AU}^3.$$

Since M is much larger than m,

$$(M+m) \approx M = 1\,\mathrm{M}_\odot$$

(where ≈ means approximately equal to).

Equation 7.1 therefore becomes

$$1\,\mathrm{yr}^2 = k \times \frac{1\,\mathrm{AU}^3}{1\,\mathrm{M}}.$$

So the constant $k = 1\,\mathrm{yr}^2\,\mathrm{M}_\odot\,\mathrm{AU}^{-3}$.

In the Solar System, the central body is always the Sun so Equation 7.1 reduces to $P^2 \approx a^3$ when P is measured in years and a in AU. For other planetary systems (see Section 7.6) the mass of the central star must be used for M in Equation 7.1. Similarly, the mass of a planet can be calculated from the orbit of a satellite (assuming it has negligible mass compared with the planet).

Just as the Earth has its Moon, some of the other planets are also orbited by smaller objects. Collectively these 'moons' are known as **satellites** – a term that has become associated with artificial objects but which originally meant natural objects.

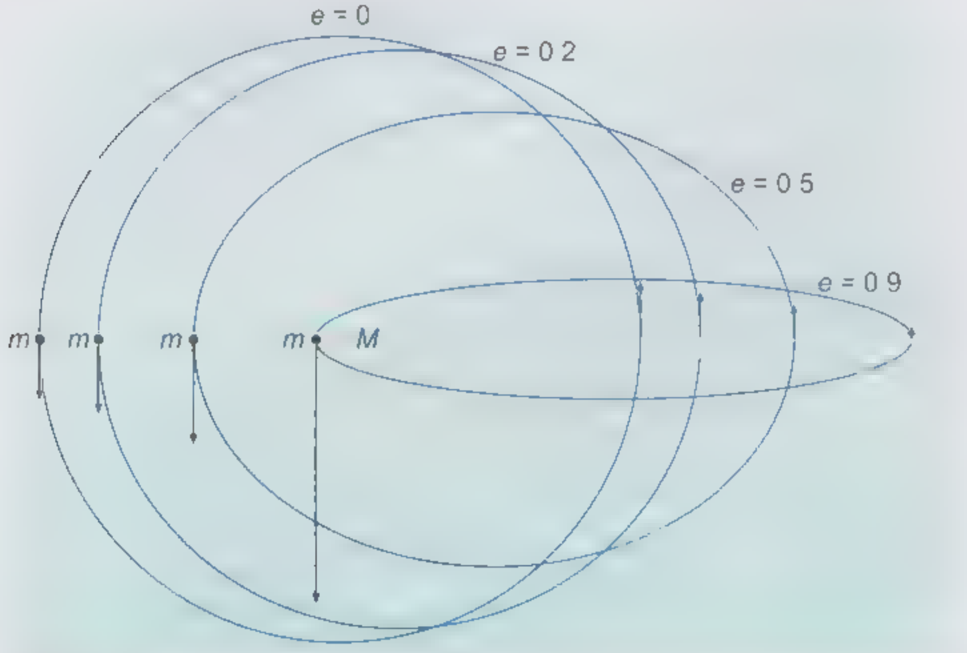


Figure 7.1 Elliptical orbits with the same average distance but different eccentricities. Note that the massive body (*M*) is not at the centre of the ellipse. The smaller body (*m*) has its highest relative speed at the point of closest approach to *M*. The orbital speed is proportional to the length of the arrow in each case.

Figure 7.2 (overleaf) shows a plan view of the Solar System that gives a better indication of the shapes and relative sizes of the orbits of the planets. Most are indistinguishable from circles. All the planets orbit in the same direction, called **prograde,** which is defined as the anticlockwise direction as seen from above (north) of the plane of the Earth's orbit about the Sun, called the **ecliptic**. In addition they orbit in almost the same plane. The tilt of the orbital plane of an object to the ecliptic is called its **inclination**, which can vary from 0° to 180°. Inclinations of greater than 90° mean the object orbits in the opposite direction, or **retrograde** (clockwise as viewed from above). Figure 7.2 shows the very low inclinations of the planets, whereas some dwarf planets and comets in particular, can have very high inclinations.

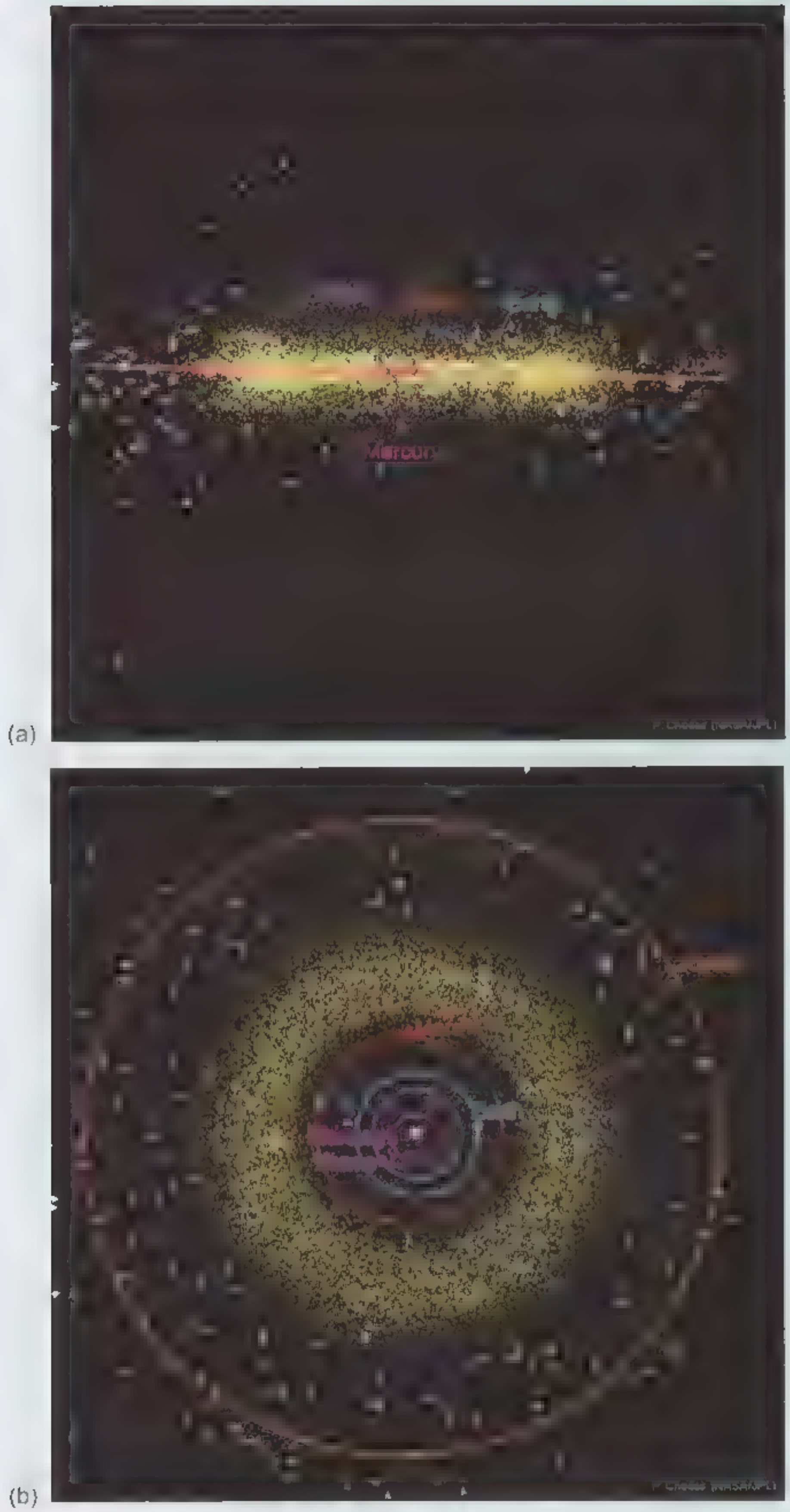


Figure 7.2 a+b Schematic views of the inner Solar System showing the orbits of the major planets and the positions of comets (white arrows) and other small bodies (yellow dots) on 1 October 2011: (a) side-on view, (b) 'top' view.

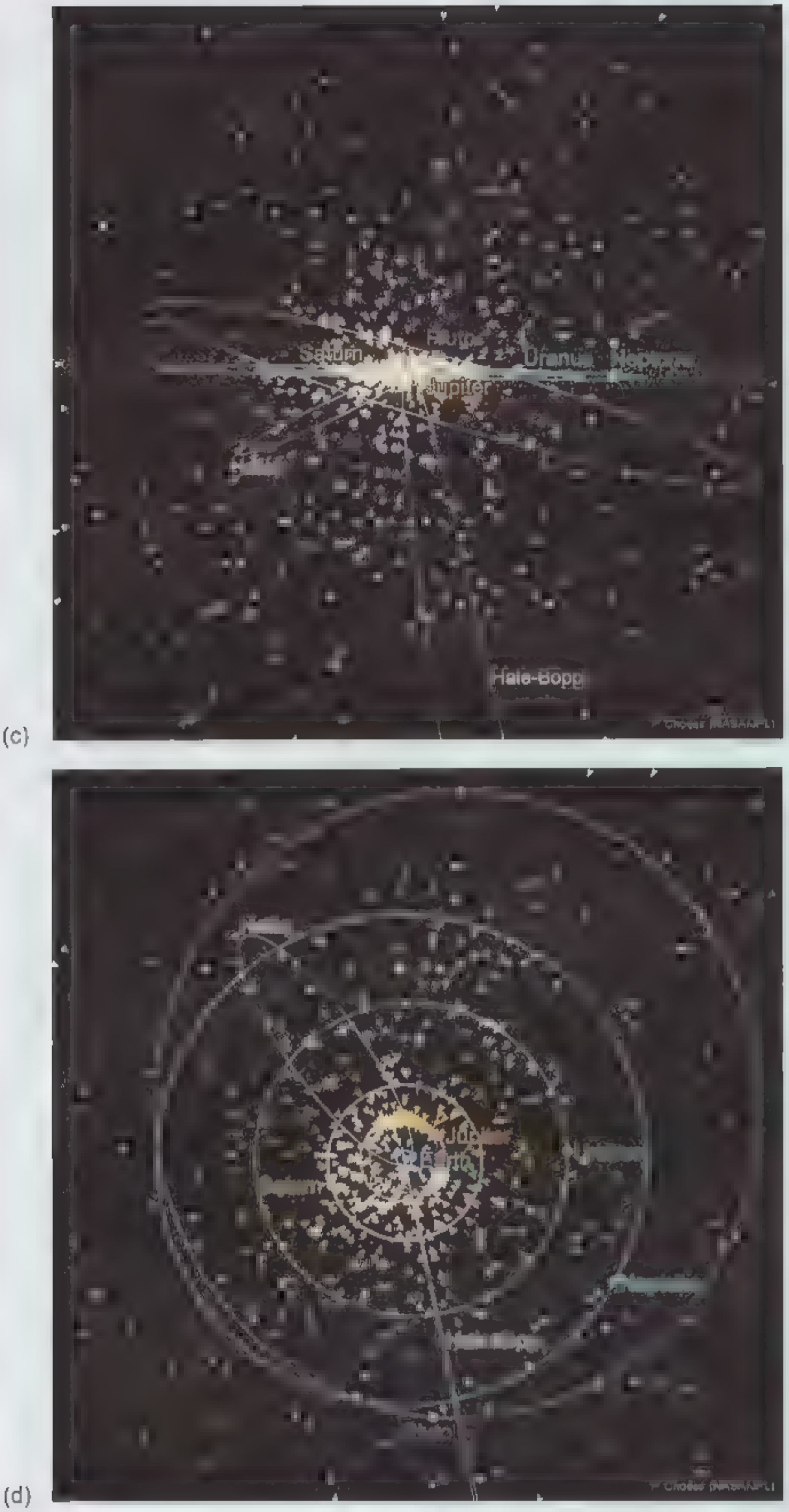


Figure 7.2 c+d Schematic views of the outer Solar System showing the orbits of the major planets, the dwarf planet Pluto and two well-known comets as well as the positions of comets (white arrows) and other small bodies (yellow dots) on 1 October 2011: (c) side-on view, (d) 'top' view.

The orbital properties of the eight planets and five dwarf planets are listed in Table 7.1. In 2006, the International Astronomical Union, the professional society of astronomers responsible for naming celestial objects, defined a new type of object, called a **dwarf planet**. A dwarf planet is a celestial body orbiting the Sun that is massive enough to be spherical as a result of its own gravity but has not cleared its neighbouring region of other objects and is not a satellite (moon). A **planet** (within our Solar System) was therefore defined as a celestial body orbiting the Sun that is massive enough to be spherical as a result of its own gravity, has cleared its neighbouring region of other objects and is not a satellite.

Table 7.1 Orbital properties of Solar System bodies.

Object	Class	Average distance /AU	Orbital period/ years	Eccentricity	Inclination[a]	Obliquity[b]	Rotation period/days
Mercury	planet	0.39	0.24	0.21	7°	0°	59
Venus	planet	0.72	0.62	0.01	3°	177°	243
Earth	planet	1.00	1.00	0.02	0°	24°	0.997
Mars	planet	1.52	1.88	0.09	2°	25°	1.03
Ceres	dwarf planet	2.77	4.60	0.08	11°	3°	0.38
Jupiter	planet	5.20	11.9	0.05	1°	3°	0.41
Saturn	planet	9.54	29.4	0.05	3°	27°	0.44
Uranus	planet	19.19	84.0	0.05	1°	98°	0.72
Neptune	planet	30.07	165	0.01	2°	30°	0.67
Pluto	dwarf planet	39.48	248	0.25	17°	120°	6.4
Haumea	dwarf planet	43.0	282	0.20	28°		0.16
Makemake	dwarf planet	45.4	306	0.16	29°		0.32
Eris	dwarf planet	68.1	561	0.44	44°		1.08
	NEAs	0.5 – 4		0 – 0.9	0° – 70°		
	Asteroid belt	1.8 – 4.2		0 – 0.4	0° – 40°		
	Jupiter Trojan asteroids	5.20		0 – 0.1	0° – 30°		
	Comets	> 2		any	any		
	Kuiper Belt	39 – 48		0 – 0.3	0° – 10°		
	Scattered Disc	30 to >500		0 – 0.9	0° – 50°		

(a) If the inclination is more than 90° the object orbits in a retrograde (clockwise as viewed from the North) direction

(b) If the obliquity is more than 90° the object rotates in a retrograde (clockwise as viewed from the North) direction.

These definitions came about because of the discovery of a number of **trans-Neptunian objects (TNOs**, objects with average distances from the Sun larger than the orbital radius of Neptune) comparable in size to Pluto and the likelihood that many new 'planets' would be found. Many TNOs have low-eccentricity, low-inclination orbits in a region known as the **Kuiper Belt**, a stable region of the outer Solar System beyond the orbit of Neptune. The **scattered disc** is a separate class of trans-Neptunian objects, with higher eccentricities and inclinations, that appear to have been scattered by Neptune (see Section 7.5).

The lower part of Table 7.1 gives the range of orbital properties for several different classes of minor bodies. The majority of asteroids are found in stable orbits in the **asteroid belt**, between the orbits of Mars and Jupiter. The dwarf planet Ceres lies in the asteroid belt and was previously classified as an asteroid. **Near Earth asteroids** (NEAs) are so-called because they have orbits which bring them close to the Earth's orbit and they are the largest objects which could potentially collide with the Earth. **Trojan asteroids** (more correctly called Jupiter Trojan asteroids) have the same orbit and period as Jupiter but lie at stable positions 60° in front of or behind the planet in its orbit about the Sun. Similar groups have been found for Neptune and also for large moons of the giant planets. Comets also have orbits which cross those of the planets but have very different properties and origins from asteroids. The properties of the orbits of all these objects provide clues to the formation and evolution of the Solar System, described in Section 7.5.

7.2.2 Spin, obliquity and seasons

An important aspect of everyday life on Earth is the change in daylight hours and consequent changes in the average daily temperature that characterise the different seasons.

- Could the seasons be caused by the Earth's orbit not being circular (i.e. the Earth is further away from the Sun in winter)?

  No. There are three reasons. (a) The Earth's orbit is only very slightly non-circular, so the effect of changing distance from the Sun is negligible. (b) Winter in the Northern Hemisphere occurs at the time of summer in the Southern Hemisphere. (c) It does not explain why the hours of daylight change with the seasons.

The seasons are caused by the Earth's axis of rotation *not* being perpendicular to the plane of its orbit but being *inclined* to the perpendicular at an angle of approximately 23.5° (see Figure 7.3 overleaf). The angle between the rotation axis and the perpendicular to the ecliptic is called the **obliquity**. The direction of the Earth's rotation axis remains fixed with respect to the distant stars but *not* with respect to the Sun. Figure 7.3 shows how the North Pole is inclined towards the Sun in the northern summer, so that in the Northern Hemisphere the Sun reaches higher in the sky, remains above the horizon longer and therefore provides more heat than in the winter.

The rotation period (approximately equal to the length of the day) and the obliquity of a planet have an important influence on the range of temperatures that the planet may experience.

- Why is the rotation period only approximately equal to the length of one day?

  The rotation period is defined as the time taken for a planet to spin on its axis once with respect to the distant stars, whereas the length of a day is the time taken for it to spin once with respect to the Sun. Because the planet moves through a small part of its orbit during a day, the direction of the Sun appears to change with respect to the background stars. The Earth moves approximately 1° around its orbit in a day so it must spin 1° more to complete one rotation with respect to the Sun. The average length of a day is defined as 24 hours, so the Earth's rotation period, defined with respect to the stars is about 359/360 as long, or about 23 hours 56 minutes.

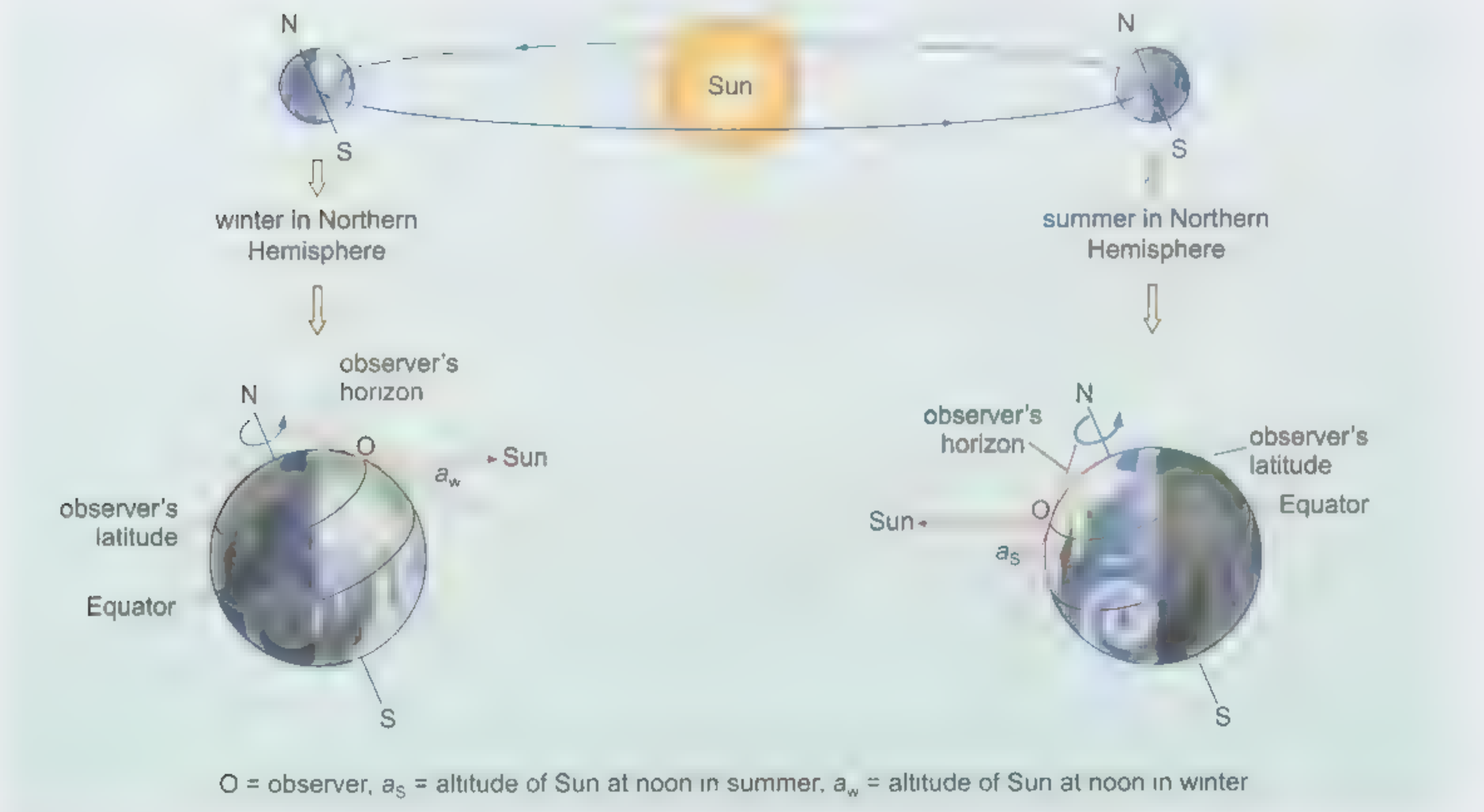

N
Sun
N
S
S
winter in Northern Hemisphere
summer in Northern Hemisphere
observer's horizon
N
O
a_w
Sun
observer's latitude
Equator
S
observer's horizon
N
observer's latitude
Equator
Sun
O
a_S
S
O = observer, a_S = altitude of Sun at noon in summer, a_w = altitude of Sun at noon in winter

Figure 7.3 The alternating seasons as the Earth orbits the Sun (not to scale).

7.2.3 Inventory of the Solar System

At the time of writing, the known members of the Solar System comprise eight planets with 174 confirmed moons, 5 dwarf planets, over half a million asteroids, three thousand comets and more than a thousand trans-Neptunian objects It is possible to estimate the total number of minor bodies in the Solar System from the properties of those that have been discovered. There are estimated to be about a million asteroids larger than 1 km, and many more at smaller sizes. The trans-Neptunian region contains around 100 times more objects than the asteroid belt. There may be several thousand undiscovered dwarf planets in the trans-Neptunian region as well as a number that have been observed but whose sizes are not yet well determined. Ceres is currently the only dwarf planet in the asteroid belt, but the definition of a dwarf planet could apply to at least three more Table 7.2 lists the physical properties of planets and dwarf planets, which we will study further in the next section.

Table 7.2 Physical properties of planets and dwarf planets.

Object	Class	Average diameter/ km	Average density/ kg m^{-3}	Atmosphere[a]	Average surface temperature[b]/K	Satellites[c]		Rings	Magnetic field strength
						Major	Minor		(Earth's field = 1)
Mercury	planet	4880	5430		440	0	0		0.01
Venus	planet	12 100	5240	dense CO_2	730	0	0		0
Earth	planet	12 700	5510	N_2, O_2	290	1	0		1
Mars	planet	6780	3930	tenuous CO_2	220	0	2		0.001
Ceres	dwarf planet	≈950	2080		170	0	0?		
Jupiter	planet	139 800	1330	dense H_2, He	(165)[b]	4	≥ 59	Yes	20 000
Saturn	planet	116 500	690	dense H_2, He	(135)[b]	5	≥ 57	Yes	600
Uranus	planet	50 700	1270	dense H_2, He	(75)[b]	4	≥ 23	Yes	50
Neptune	planet	49 200	1640	dense H_2, He	(70)[b]	1	≥ 12	Yes	30
Pluto	dwarf planet	2300	2050		≈45	1	≥ 2		
Haumea	dwarf planet	≈1200	≈3000		≈40	0	≥ 2		
Makemake	dwarf planet	≈1500			≈40	0	0?		
Eris	dwarf planet	≈2400	≈2400		≈35	0	≥ 1		

(a) CO_2 is carbon dioxide, N_2 is nitrogen gas (it has two N atoms in each gas molecule), O_2 is oxygen gas (two O atoms per gas molecule), H_2 is hydrogen gas (two H atoms per gas molecule) and He is helium gas (a monatomic gas, so only single helium atoms).

(b) The average surface temperatures for the gas giants are defined at a depth where the atmospheric pressure is equal to that at the surface of the Earth.

(c) Major satellites are defined as having a diameter of 800 km or more.

7.3 Planets

The individual planets have fascinated human beings for centuries. Unlike the stars, which appear in fixed patterns night after night, the planets move across the sky. Once it was established (in the fifteenth and sixteenth centuries) that the Earth and other planets all orbit the Sun, and the development of telescopes gave astronomers a more detailed view, people began to wonder about life on other planets. With improvements in observational techniques and the advent of space exploration, it's now known that intelligent life is not present elsewhere in the Solar System but microbial life could exist and is still being sought (see Chapter 8).

7.3.1 Classifying planets

Our current knowledge of the planets and their satellites is based on information collected by space probes as well as telescopes on the ground and in space. Together with knowledge gained from studying planet Earth, such observations give an understanding of conditions on other planets, and some clues about how the Solar System might have formed.

One way in which astronomers (and scientists in general) try to make sense of what they observe is to look for patterns and similarities. Astronomers have looked for ways in which planets resemble one another so they can talk about groups of planets rather than individual ones. This has helped shed some light on their history. The next activity concerns this classification process.

Activity 7.1 Classifying planets

The estimated time for this activity is 30 minutes

You will find at least one image of each of the major planets in this book. There are many spectacular images available from the space missions that have visited the planets over the last fifty years. The notes for this activity (in the Activities section of the module website) also contain a small collection of images. You may wish to visit some of the websites, from which these images originated, that are also listed on the module website. Study these images and their captions. Tables 7.1 and 7.2 list the basic properties of each planet.

Suggest ways in which planets might be classified on the basis of this information. What characteristics do you think would lead to a helpful classification system? Size? Colour? Presence of rings? Number of natural satellites? Some other characteristic? For each characteristic, try to think of reasons why it might, or might not, be a sensible way to classify the planets.

Several of the characteristics suggested in Activity 7.1 (size, satellites and rings) lead to the same way of grouping the planets. Astronomers generally classify the planets into two main groups. The **terrestrial** (Earth-like) **planets** – Mercury, Venus, Earth and Mars – lie closest to the Sun, are fairly similar in size, and have few or no natural satellites (two at the most) and no rings. The **giant planets** (sometimes referred to as gas giants) – Jupiter, Saturn, Uranus and Neptune – are very much larger than the terrestrial planets, and lie much further from the Sun. Their orbits are also much more widely spaced than those of the terrestrial planets. Each giant planet has several satellites and some rings, with those of Saturn being by far the most prominent.

The broad similarities within the two main groups of planets do not relate just to their appearance but also to their interior structures.

7.3.2 The terrestrial planets

The terrestrial planets are made from rocky materials and have structures similar to that of Venus shown in Figure 7.4. They have dense cores mainly of iron, almost everything else being **silicate** rocks (composed of compounds containing silicon and oxygen). The outermost layer behaves rigidly and is called the **lithosphere** (meaning 'rocky shell'). Between this layer and the core lies the mantle, where the silicate material is so hot that, even though it is not molten, it can flow at a rate of a few centimetres per year. This weak interior is stirred up by convection currents that transport heat towards the surface.

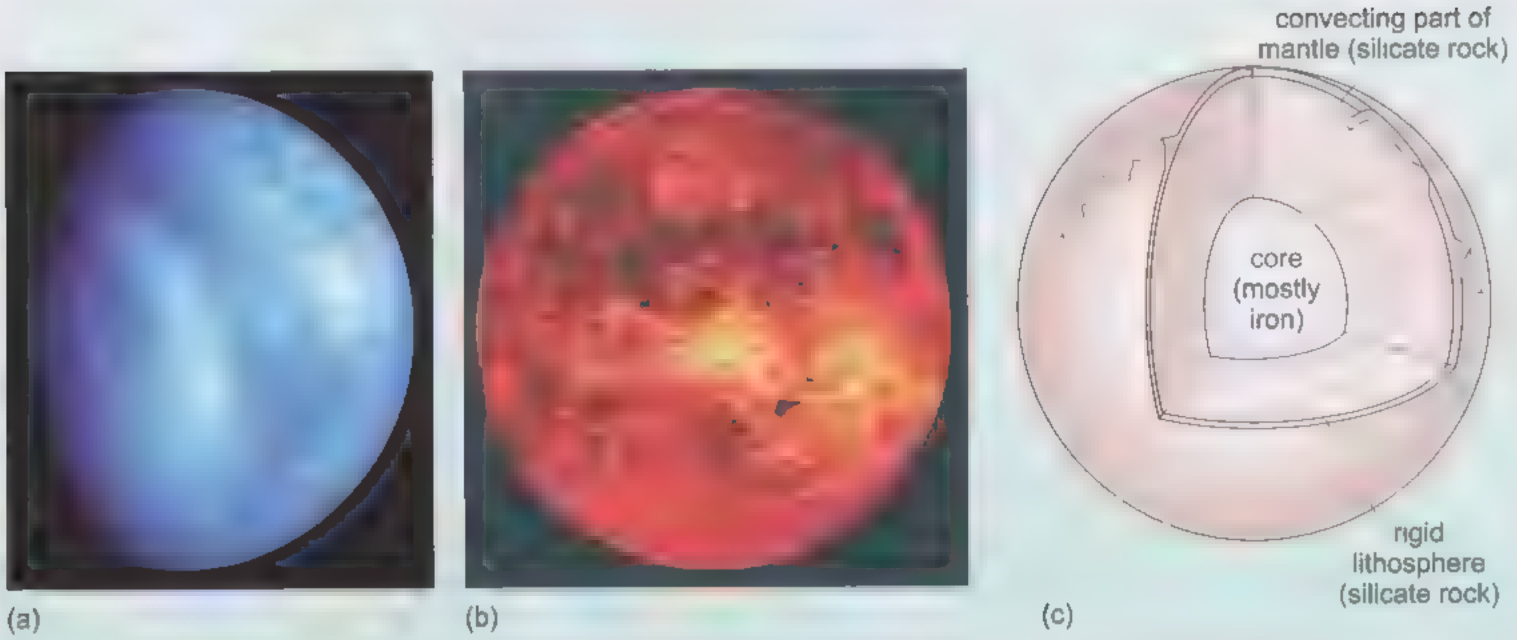

convecting part of mantle (silicate rock)
core (mostly iron)
rigid lithosphere (silicate rock)
(a)
(b)
(c)

Figure 7.4 (a) The planet Venus imaged by the Venus Express spacecraft in 2007 in ultraviolet light. The blue appearance is false colour used to enhance the structure of the clouds, the clouds are uniform and white when viewed through a telescope. (b) Image of Venus made using radar data taken by the Magellan spacecraft in 1990 to penetrate the dense atmosphere. (c) Cut-away view of the interior of Venus.

What you see when you look at a solid planetary body depends on whether or not there is enough heat escaping from the interior so that the lithosphere remains thin enough for it to be punctured by volcanism and deformed by the underlying convection currents. In the case of the Earth, these processes have driven the motion of large 'plates' that make up the surface and created the continents and ocean basins and most of the mountain ranges on the Earth (such as the Alps, the Himalaya and the Andes).

Planets with lithospheres that have grown too thick for these processes to be effective show progressively less evidence of volcanism and deformation; these surface traces gradually become obliterated by craters from the occasional impacts of asteroids or comets that continue even today.

Although the atmospheres of the terrestrial planets have insignificant mass compared with the mass of the planet, they play an important role in shaping the environment on the planet's surface. Mercury has essentially no atmosphere and suffers extreme ranges of temperature from around 740 K at the equator around midday, down to 80 K after the long Mercurian night. Despite its very long day, the temperature range on Venus is very small due to its dense atmosphere, but despite its greater distance from the Sun it is no less hostile than Mercury, with average temperatures of 730 K and atmospheric pressure ninety times that of Earth!

The Earth has a magnetic field that is similar in its properties to a magnetic dipole (it acts much like a bar magnet through the Earth), but it is not perfectly aligned with the rotation axis. This field results from the dynamo formed by the motion of an electrically conductive liquid outer core (known to exist from the measured properties of seismic waves). The dipole effect is believed to be due to currents generated in this inner core by interactions between the Earth's rotation and convection.

■ Why does Venus, which is similar in size to the Earth and with a similar sized iron core, not have a dipole magnetic field like the Earth?

□ Venus rotates extremely slowly compared with the Earth. In addition, it may also lack a solid inner core, which may also be necessary to heat the surrounding liquid to create convection.

The influence of atmospheres and magnetic fields on the potential for the existence of life in different planetary environments is explored in Chapter 8.

7.3.3 The giant planets

Jupiter and Saturn are made largely of hydrogen and helium (Figure 7.5). This is directly visible in their atmospheres, but knowledge of how materials behave, deduced in Earth-based laboratories, leads astronomers to conclude that the atmospheres of these two planets become gradually thicker with depth. The high pressures cause hydrogen atoms to be squeezed so close together that they behave like liquid metal. The core at the centre is believed to consist largely of water and other ices and rocky materials, which may be liquid at the high temperatures that occur there.

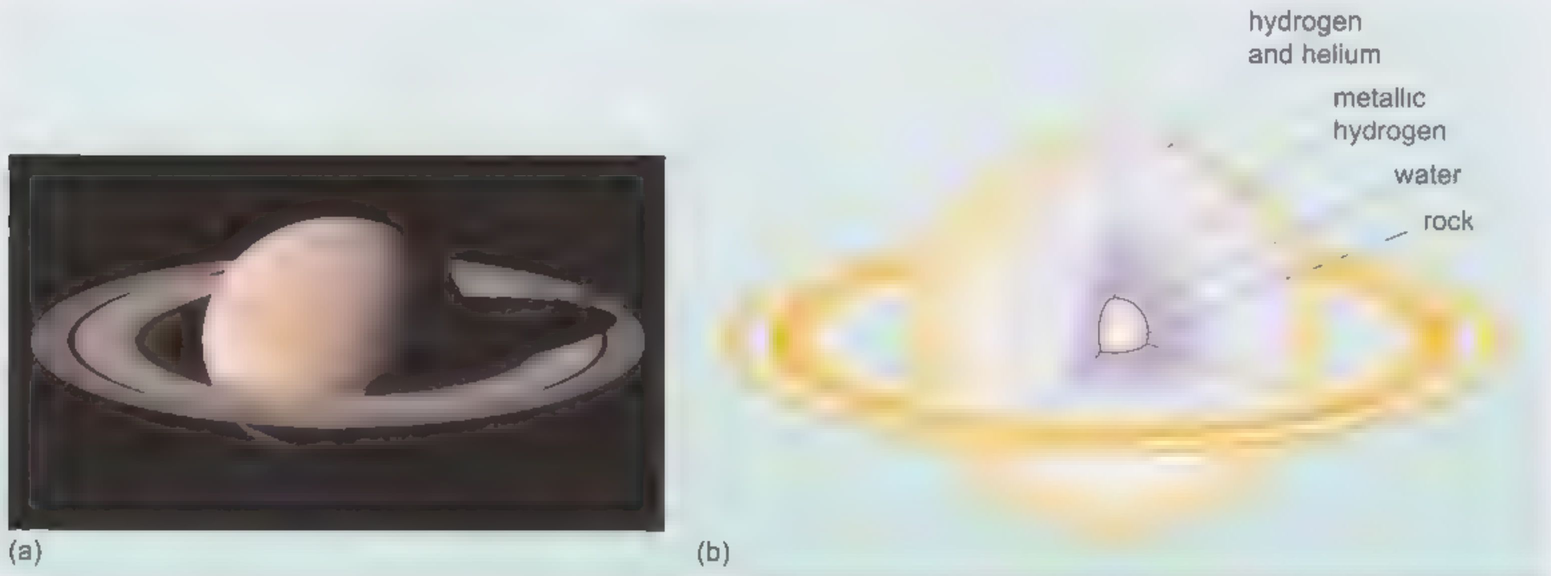


Figure 7.5 (a) Image of the planet Saturn taken by the Cassini spacecraft in 2004. (b) Cut-away view of the interior of Saturn.

Uranus and Neptune can be thought of as resembling Jupiter and Saturn, but with less massive envelopes of hydrogen and helium, so they contain a substantially higher proportion of water and rocky materials.

7.3.4 Icy bodies

Many objects in the outer Solar System, such as satellites of the gas giants or trans-Neptunian objects, are reckoned to contain large amounts of water ice, based on their densities and the spectra of their surfaces.

- Two sorts of materials that make up planets can be broadly classified as either **icy materials** or **rocky materials** Suggest definitions for these two terms.

  Rocky materials are solid at temperatures typical of the Earth's surface and include rocks (!), soil and metals. Rocky materials melt at high temperatures, such as occur naturally deep inside planets (or are created artificially in furnaces). Icy materials are normally liquid or gas, or else they melt very easily, at the Earth's surface. Icy materials include water (in the form of ice, liquid water or water vapour), carbon dioxide, ammonia and methane. At sufficiently low temperatures (such as those at the surfaces of planets lying far from the Sun), icy materials are solid.

An icy body in the outer Solar System can behave in a similar way to a rocky body such as a terrestrial planet. They may have cores of rocky material, but are deeply overlaid by ice. The outermost layer of ice is so cold that it is very rigid and acts in the same way as rock on Earth, forming an icy lithosphere. Below this, provided there is a supply of internal heat, the ice becomes mobile enough for convection, leading to the same range of surface deformation and volcanism that is displayed on the terrestrial planets, except that the volcanism involves melts derived from ice rather than molten rock. In some cases there may be a reservoir of liquid ice below the lithosphere. This is considered further in Chapter 8.

7.4 Small Solar System bodies

7.4.1 Asteroids

Asteroids are rocky objects mostly orbiting in the asteroid belt, which is found between Mars and Jupiter. Their total mass is much less than that of a terrestrial planet and the largest, Ceres (now classified as a dwarf planet) contains almost one-third of the mass of the belt. Many asteroids are believed to contain rocks that have been relatively unaltered since their formation. Therefore, they may be typical of the material from which the terrestrial planets formed.

Asteroids produced many of the impact craters on the terrestrial planets and many have themselves been heavily bombarded (Figure 7.6). Many other asteroids have had catastrophic collisions, producing groups of fragments with similar orbits called *asteroid families*. Astronomers believe that many of these bodies consist of re-accumulated material with little internal strength called **rubble piles**.

■ How would the density of a rubble pile compare with the density of the pre-impact body from which it originated?

□ A rubble pile would have some empty gaps between its constituent fragments, so it would have a lower density than a solid rock of the same size.

One of the few asteroids with a measured density, Mathilde (Figure 7.6), has a density of only $1300\ \text{kg m}^{-3}$, (compared with a density for the solid rocks of which it is composed of around $2800\ \text{kg m}^{-3}$) implying that it is a rubble pile with more than 50% void spaces in its interior.

The composition of asteroids, like planets, depends on how and where they formed and how they have been processed by heating, collisions and chemical processing since their formation. Asteroids are classified according to the way they reflect sunlight. These *reflectance spectra* of asteroids do not contain narrow spectral lines (as found from the gases in the atmospheres of stars) but do have some characteristic broad features which result from differences in surface composition. There are many types, but the most common are listed in Table 7.3.

Table 7.3 Selected asteroid types.

Asteroid type	Location	Albedo[a]	Spectrum	Inferred composition	Meteorite type (see Section 7.4.3)
C	more common in outer asteroid belt	low (< 10%)	neutral	carbon-rich, rocky	carbonaceous chondrite
S	more common in inner asteroid belt	moderate (10 to 25%)	red, broad absorptions due to silicate minerals	silicates or silicate–metal	stony or stony–iron
M		moderate (10 to 20%)	neutral	metallic	iron

[a] Fraction of light reflected.

7.4.2 Comets

Comets are typically a few kilometres across and consist largely of water ice and rocky particles (Figure 7.7a). They can have extremely elongated orbits, inclined at any angle to the plane of the Solar System (Figure 7.2). When they stray into the inner Solar System, some of the ice vaporises, dragging dust particles with it, giving rise to enormous tails and sometimes spectacular sights in the night sky (Figure 7.7b).

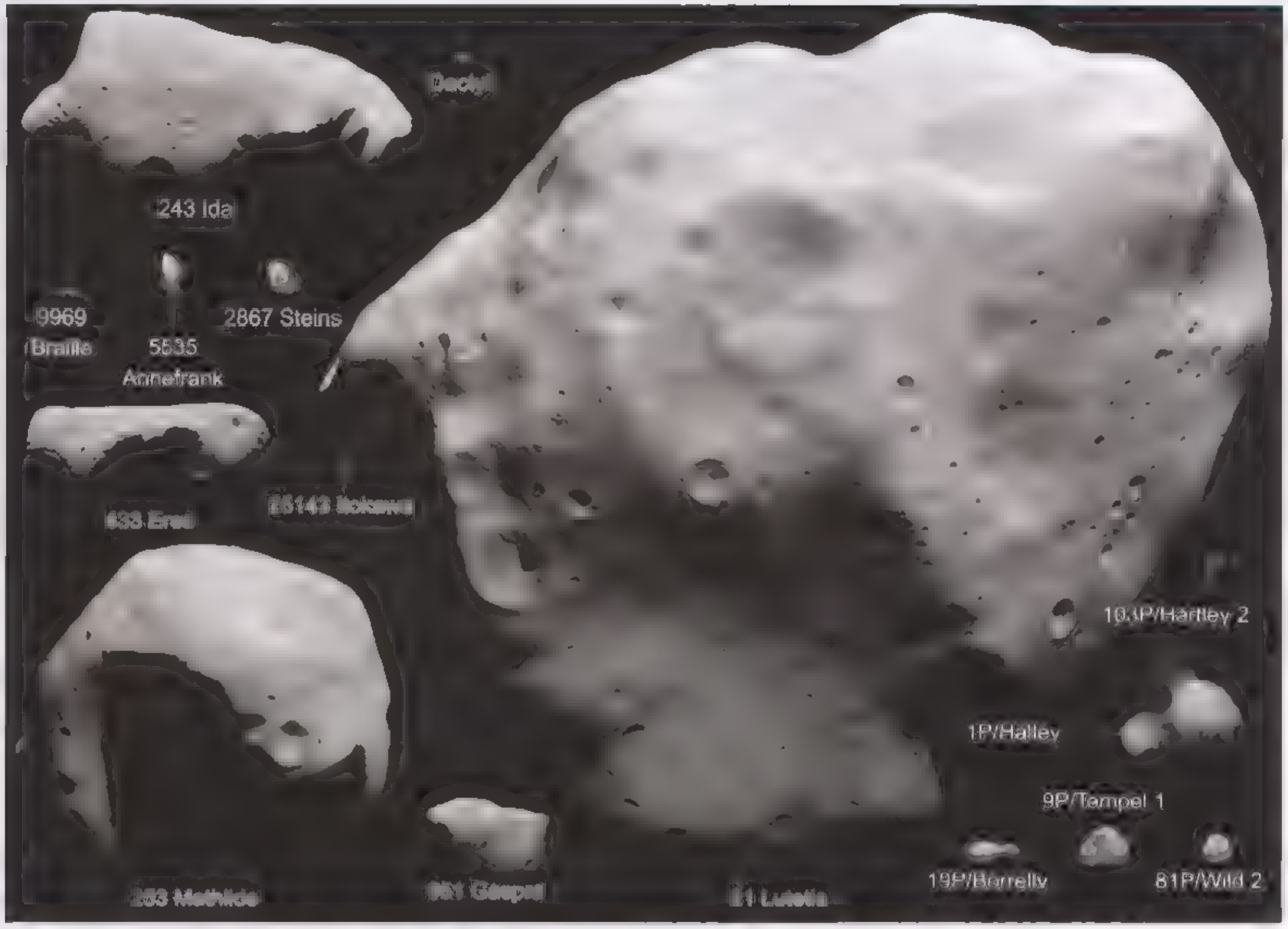


Figure 7.6 All small Solar System bodies visited by spacecraft (as at beginning of 2011). The objects are shown to scale; the largest, Lutetia, has a diameter of 120 km along its longest axis and the smallest, the near-Earth asteroid Itokawa, is less than a kilometre in diameter.

The icy composition of comets suggests an origin in the outer Solar System. Most of them are currently in two regions:

i in a spherical cloud (the **Oort cloud**) surrounding the Solar System, extending perhaps one-third of the way to the nearest star, and containing icy bodies thrown out after close encounters with the newly forming Jupiter; and

ii trans-Neptunian objects in the outer Solar System (predominantly those in the scattered disc) centred on the plane of the Solar System.

The gravitational influence of passing stars or close encounters (with each other or with the outer planets) occasionally causes an object from the Oort cloud or trans-Neptunian region to enter the inner Solar System and develop a visible tail. These **long period comets** pass briefly through the inner Solar System in highly eccentric orbits with periods of hundreds to millions of

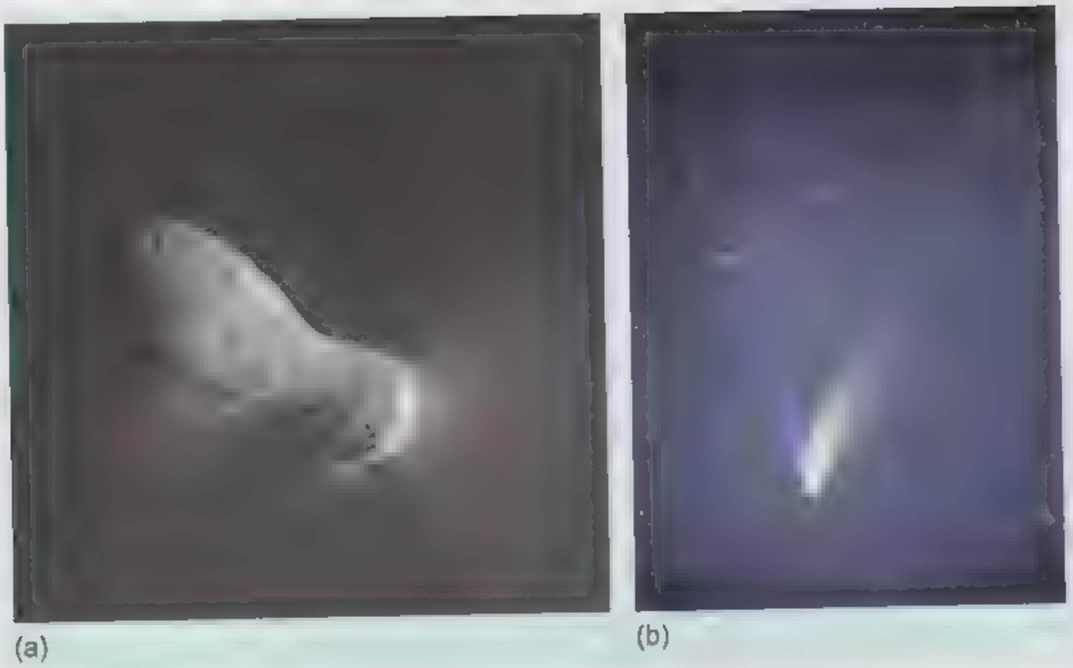


Figure 7.7 (a) The 2 km long nucleus of comet Hartley 2 imaged by the EPOXI (formerly Deep Impact) spacecraft in 2010. (b) Comet Hale–Bopp, which was visible to the unaided eye in 1997; the tenuous comet tail extends many millions of kilometres. The nucleus is far too small to be visible in this image.

years. If they encounter a major planet, they may be captured into **short period comet** orbits (periods less than 200 years) where they may survive for a few thousand orbits before being ejected by another close encounter or exhausting the ices near their surface and ceasing activity.

7.4.3 Meteors and meteorites

Extraterrestrial fragments of material entering the Earth's atmosphere are sometimes observed as 'shooting stars' when they heat up on passing through the atmosphere — more precisely, they are called **meteors**. Although meteors can be seen at any time and from any direction (a few per hour from a dark site), many arrive in *meteor showers*, when rates can reach as high as one per minute for a short while. Meteor showers occur at the same time each year as the Earth crosses the orbit of the parent comet and intercepts dust particles emitted by the comet during previous passages close to the Sun.

Larger objects that survive passage through the atmosphere and reach the Earth's surface are called **meteorites**. Most are pieces of asteroids but a few are fragments chipped off the surface of the Moon or Mars (where the low gravity and tenuous atmosphere enable them to escape).

■ Meteors, particularly those that arrive in showers, originate from comets. Why do we not appear to have meteorites from comets?

Comets are icy bodies that formed in the outer Solar System. The ices cannot survive the temperatures in the inner Solar System and the remaining solid particles are too fragile to survive atmospheric entry except as very fine dust (giving rise to micrometeorites).

Meteorites provide an invaluable source of extraterrestrial material that can be analysed in microscopic detail in laboratories on Earth to provide information on the conditions in the early Solar System, when the rocky material first solidified, and about the conditions on the asteroids they formed. Meteorites are classified according to their composition and structure.

Iron meteorites are predominantly composed of iron metal, while stony meteorites are composed mostly of silicate minerals. Stony iron meteorites are mixtures of metals and silicates. There are many types of stony meteorites, some of which contain minerals that are either wholly or partly melted and then recrystallised. The most primitive (i.e. least altered since their formation), are called carbonaceous chondrites. They are carbon-rich meteorites that contain organic compounds including amino acids as well as chondrules (small silicate spheroids that were formed before being incorporated into meteorites). The compositions of the main asteroid types were inferred from comparisons of the spectra with known meteorites (see Table 7.3).

Martian meteorites contain tiny amounts of trapped gas, which tell us about the state of the Martian atmosphere at the time the rocks solidified. Even though we have lunar samples from the NASA Apollo and Soviet Luna missions, lunar meteorites provide rocks from different regions of the Moon.

7.5 A brief history of the Solar System

7.5.1 Formation of the planets

Theories of how the Solar System formed can account for the differences in composition among the various planets; broadly speaking, the controlling factor is distance from the Sun. Although many details are still not fully understood, the Solar System appears to have formed within a rotating disc of gas and dust known as the **solar nebula** nearly 4600 million years ago. The initial composition of the solar nebula was that of the dense cloud core from which it formed, with hydrogen, helium and a wide range of molecules as well as silicate and carbon-rich dust grains, with mantles (i.e. coverings) of water and other ices.

As the cloud collapsed, the rate of rotation of the disc increased due to the *conservation of angular momentum* (the same effect that causes a skater to spin faster as he or she draws arms inwards, see also Activity 6.1). The central part of the nebula was heated to high temperatures by the release of gravitational energy, vaporising the original ices and dust. As the disc lost mass to the newly forming star at its centre, the density and temperature in the disc started to fall. In the inner part of the disc, the only substances to condense in large amounts were metals and silicate minerals that make up rock. Further out, it was cold enough for water, as well as other volatile (i.e. low melting point) ices such as ammonia and methane, to condense. The **snow line** is the spherical boundary in the solar nebula outside which it was cool enough for water to condense into solid ice grains.

Because all the condensed dust grains were travelling in nearly circular obits in the same plane they collided at very low speeds, building aggregates of larger and larger size. Once these aggregates reached a size of around

10 kilometres they began to attract each other by gravity. Computer models simulating the growth of these **planetesimals** show that they rapidly form a few large bodies, **planetary embryos**, at different distances from the Sun. The final stages of planetary formation, due to collisions (including possible fragmentation and re-aggregation, as is believed to have formed the Earth Moon system) of planetary embryos to form **protoplanets**, occurred more slowly, particularly at larger distances from the Sun. These protoplanets then swept up the remaining planetesimals and nebular gas and dust to form the planets we see today. The protoplanets close to the Sun are predominantly made of the high-temperature metals and minerals that condensed there. The protoplanets that formed outside the snow line grew so big that they were able to capture substantial amounts of the gaseous hydrogen and helium that made up most of the nebula. These became the giant planets.

■ Why could the protoplanets beyond the snow line grow so large?

Hydrogen and oxygen are very common gases so large amounts of water ice can condense beyond the snow line as well as the rocky dust grains. As the distance from the centre of the nebula increases, so the volume of space around the orbit of a planet, and hence the mass of planetesimals within it, increases.

An alternative theory for giant-planet formation is that the planets condensed directly from the nebula, rather than by capture of nebular gases by large icy protoplanets. In this theory the planets initially had a composition like the Sun, but accreted rocky and icy planetesimals, resulting in giant planets with similar composition.

Some of the icy and rocky material that gathered in the vicinity of each giant planet appears to have avoided capture onto the planets themselves and went instead to form their satellites.

In the outer reaches of the solar nebula, beyond the orbit of Neptune, the planet-forming process did not progress to form a giant planet, possibly because there was not enough material in the nebula at these distances or because the process was interrupted by the influence of newly-formed planets closer to the Sun.

7.5.2 Dynamical evolution

After the most massive planet Jupiter formed, its gravity played a major role in the evolution of both comets and asteroids. Jupiter's gravity disturbed the orbits of nearby planetesimals sufficiently for collisions to cause fragmentation rather than accretion. The current asteroid belt is the remnant of the original planetesimals and planetary embryos that were prevented from accreting to form a planet. Only a few original large planetesimals, such as Ceres and Vesta, remain, with most asteroids being the result of one or more collisional fragmentations. The majority of meteorites are small fragments from these collisions.

The newly formed Jupiter also had a profound influence on icy planetesimals. Those that approached the planet but were not accreted were scattered into new orbits. They were then either accreted by other planets or the Sun or were ejected on highly eccentric orbits, to form the Oort cloud.

In the last twenty years the idea that planets formed at more-or-less their current distances from the Sun has been questioned. This was not just the result of observations of the structure and composition of the Solar System but also the discovery of planets around other stars. Many of these planets, termed **hot Jupiters** (see Section 7 7), appeared to have properties similar to Jupiter but have orbits very close to their parent stars.

■ Why cannot a planet the size of Jupiter form very close to a star?

There would be insufficient material in the nebula to make the planet and the temperature would be too high to allow the condensation of icy or most rocky materials.

The distribution of the orbits of the trans-Neptunian objects, including many in Pluto-like orbits and in the scattered disc, provided clues to Jupiter's formation. Attempts to explain these observations led to the definition of the Nice model (so-called because it was developed by scientists at the Observatoire de la Côte d'Azur in Nice in 2005) for the dynamical evolution of the Solar System. Although the details are beyond the scope of this module, the theory requires that after their formation, the gas giant planets had orbits much closer together than now. The Kuiper Belt was much denser and closer to the Sun. The orbits of the planets evolved due to their mutual gravitational influences and those of the remaining planetesimals The net effects were that Jupiter moved slightly inward while the other planets moved outwards. Neptune, originally closer to the Sun than Uranus, moved outside the orbit of Uranus and disrupted the original Kuiper Belt. Many of these planetesimals were scattered inward, planet-by-planet, until interactions with Jupiter kicked them into highly eccentric orbits or even out of the Solar System. The outward migration of Neptune further disrupted the Kuiper Belt.

Although the Nice model is widely agreed to provide the best explanation for many of the current features and historical events in the Solar System, it is still not universally accepted. It successfully predicts:

- the Late Heavy Bombardment (a period of increased impact rates about 4 billion years ago),
- the low mass of the present day Kuiper Belt,
- the presence of the Oort cloud from planetesimals scattered by Jupiter,
- the orbits of the Trojan asteroids of Jupiter (and asteroids in a similar position with respect to Neptune),
- the scattered disc and the concentration of Kuiper Belt objects in certain orbits (like Pluto in a 3 : 2 resonance with Neptune, i.e. Pluto orbits the Sun twice while Neptune does three orbits).

The Nice model is less successful predicting:

- the numbers of small satellites captured by the giant planets,

- the Kuiper Belt objects in low-inclination orbits which appear to have different spectra and therefore an inferred different origin from the scattered disc.

7.5.3 Physical and chemical evolution

The appearance of bodies in the Solar System today is a result of the conditions prevailing when they formed combined with any evolutionary processes they have undergone since their formation. In many cases, these processes have significantly changed their properties For example, on Earth, the continual reprocessing of surface rocks through plate tectonics and erosion has mostly obliterated the record of impact craters that are seen so clearly on the Moon, Mercury and Mars. By studying different planets, we can learn more about the history of our own planet and how the whole Solar System formed and evolved.

After their formation, the interiors of the planets (as well as dwarf planets and the larger planetary satellites) underwent a process called **differentiation**. This is a process in which different constituents sorted themselves into layers of distinct composition, usually as a result of heating, where the denser materials sink towards the centre, while less dense materials rise to the surface. The details are complicated by the chemical interactions of different materials under the heating process, but the end result for a terrestrial planet is a structure with a dense core, composed mostly of the common elements iron and nickel, with a surrounding rocky mantle and surface crust The mantle is composed of rocks rich in magnesium silicates whereas the crust is composed mainly of less dense silicates of aluminium, sodium, calcium and potassium.

■ What was the source of energy for the heating that caused differentiation?

The energy comes from two main sources: the gravitational potential energy released from the aggregation of planetesimals and the decay of radioactive atomic nuclei with short half-lives. The half-life is the time taken for half the atoms of a radioactive element to decay; a short half-life means a fast decay which therefore releases large amounts of nuclear energy over a short time interval. Planets and planetary embryos were sufficiently massive that this energy could not easily be transported to their surfaces and escape and hence it heated their interiors.

There is evidence of evolutionary processes in the types of meteorites found on the Earth Iron meteorites originated in the cores of large planetesimals that had been differentiated and then catastrophically disrupted by a collision before a small fragment found its way to the Earth.

In the outer Solar System, differentiation can occur in icy bodies, which have a rocky core and icy mantle. As you will see in Chapter 8, the trapped heat can cause melting of the ices and form subsurface oceans in some bodies.

The surfaces of Solar System bodies provide evidence of any evolutionary processes that have occurred since they evolved. The number of impact craters provides an indication of the relative ages of surfaces on a body, the older the rocks, the more heavily cratered they will be. Processes such as volcanism can produce new rocks and therefore reset the cratering record The presence of an atmosphere may reduce the amount of cratering and introduces the likelihood of erosion by wind, water or ice flow.

Activity 7.2 Shaping planetary surfaces

The estimated time for this activity is 20 minutes.

In the notes for this activity in the Activities section of the module website you will find some images which illustrate processes that shape the solid surfaces of planets and satellites. As part of this activity you should look through these images to find at least one that illustrates each of the following:

- craters produced by impacts of small bodies from space;
- active volcanoes;
- mountains produced by volcanoes;
- regions shaped by rocky lava flows;
- regions shaped by icy lava flows;
- channels or canyons produced or modified by water;
- features produced by wind action.

The module website contains a tiny sample of the spectacular pictures from recent space missions. There are details of the missions, their experiments and the results on the NASA and ESA websites (links to these websites can be found on the module website).

The masses of the terrestrial planets were too low, and their temperatures too high for them to retain atmospheres of the abundant hydrogen and helium gas from the original solar nebula. The atoms or molecules in an atmosphere are retained by gravity – the larger the gravitational force at the surface of the planet, the greater its ability to retain an atmosphere. However, it is not as simple as that. All atoms and molecules in a gas are in motion and the higher the temperature of the gas and the lower the mass of the molecule, the faster the average speed of the molecules. If even a small fraction of the fastest molecules have speeds that exceed the escape velocity of the planet, then the atmosphere will gradually be lost. This is what happened to the original (primordial) atmospheres of all the terrestrial planets.

■ If the dominant atoms from the solar nebula have been lost, where have the atmospheres of the terrestrial planets come from?

□ They are the result of outgassing of volatile material (dominated by carbon dioxide (CO_2) and nitrogen (N_2)) released from the rocks in the planets' interiors.

The molecules present today in the terrestrial planets' atmospheres (see Table 7.2) are much more massive than hydrogen and helium and so their thermal speeds are much lower than the escape velocities and they can be retained over the age of the Solar System. Mercury is too small (as is the Earth's Moon) and too hot during the day to retain even these more massive molecules.

■ Assuming the rocks forming the mantles of the terrestrial planets produced similar gases through volcanic activity, why are their atmospheres so different?

□ They have evolved through chemical reactions. This evolution is highly dependent on the presence and nature of water.

Only on the Earth are conditions suitable for large reservoirs of liquid water to be present Carbon dioxide dissolves in water and is deposited as carbonate rocks, reducing the carbon dioxide content of the Earth's atmosphere. On Mars the majority of the water is solid (in the polar ice caps and below the surface) and on Venus it is all in gaseous form. The high abundance of oxygen in the Earth's atmosphere is a result of chemical conversion of carbon dioxide by plants in photosynthesis (see Chapter 8).

Understanding the properties of planets in our Solar System, how they formed and evolved, will help us to understand planetary systems detected around other stars. However, as you will discover in the next two sections, even the detection of these **extrasolar planets** has proved extremely challenging, and our knowledge of their properties is currently very limited.

7.6 The search for extrasolar planets

7.6.1 Can we see an exoplanet?

The first confirmed detection of a planet orbiting another star was made in 1992, nearly four centuries after the invention of the telescope, and after decades of use of sophisticated astronomical detectors on large telescopes from the Earth and in space. Why did it take so long? This question can be answered by considering how easy it would be to detect a large planet around a relatively nearby star.

It's possible to calculate how bright the planet would appear if you tried to observe a planet like Jupiter (the most massive in the Solar System) orbiting another star like the Sun. It turns out that the Jupiter-like planet will be sufficiently bright to be detected with a large telescope if the star it orbits is relatively close, e.g. less than 100 pc away.

■ Why may such a planet be impossible to detect directly even when it is sufficiently bright to be detectable with a large telescope?

Jupiter is visible because it reflects sunlight. A Jupiter-like planet would also reflect sunlight from its parent star. If the angular separation of the planet from the star, as seen from the Earth, is too small, the star and planet would appear as one object.

A planet, at a distance of 5 AU from its parent star 10 parsecs from the Earth will have an angular separation (see Section 2.4.2 – the angular separation is directly analogous to the angular diameter) of

$$57^\circ \times \frac{5\,\text{AU}}{10\,\text{pc}} = 57^\circ \times \frac{5\times1.5\times10^{11}\,\text{m}}{10\times3.09\times10^{16}\,\text{m}} = 57\times2.5\times10^{-6}\text{ degrees}$$

$$= 57\times60\times60\times2.5\times10^{-6}\text{ arcseconds} = 0.5\text{ arcseconds.}$$

This angular separation is distinguishable with large terrestrial telescopes using special techniques to correct for the effects of atmospheric turbulence, and easily achievable with a telescope in space. However, the star is around 100 million times brighter than the planet. This enormous brightness contrast means that the planet will be lost in the glare of the star (the result of unavoidable optical effects in the telescope as well as atmospheric scattering) and extremely difficult to detect.

So how can astronomers attempt to image planets around other stars? One approach is to try to reduce the contrast in brightness between the planet and the star so that the planet can be imaged directly. An instrument called a coronagraph (so-called because it is used to block out the disk of the Sun so that the solar corona can be observed) can be used to block the light from the star (see Figure 7.8).

In Section 2.5, you saw that cool objects emit the majority of their electromagnetic radiation at longer wavelengths than hot objects. So, while the majority of sunlight is emitted in the visible part of the spectrum, a planet like Jupiter with a temperature of less than 200 K will, as well as reflecting sunlight, emit radiation at wavelengths of tens of microns. The contrast in brightness will therefore be lower at these wavelengths, although the detectors required to observe at such wavelengths are less efficient than those used in optical telescopes. The small number of directly imaged exoplanets all have masses several times that of Jupiter and large orbital distances (from 2 to over 2000 AU).

The presence of planets can also be inferred from imaging of discs of material around stars. Many young stars have discs of gas and dust in which planets may be forming, and some older stars have dust discs that must be replenished from collisions of larger bodies (asteroids, comets and maybe planets). Cavities, clumps or warps in the discs may be caused by the gravitational influence of invisible planets. The first detections of exoplanets were made using very different techniques. The following sections provide a brief summary.

7.6.2 Detection using stellar motions

A planet is held in orbit around a star by the star's gravitational force, but the planet also exerts a gravitational force on the star. The star and planet orbit around the **centre of mass** of the system. In a system with only two bodies, the centre of mass lies on a line joining the two bodies and is closer to the more massive body. The distances from the centre of mass, r_s and r_p, and the masses, m_s and m_p, are related by

$$m_p\, r_p = m_s\, r_s \qquad (7.2)$$

where the subscripts s and p refer to the star and planet respectively.

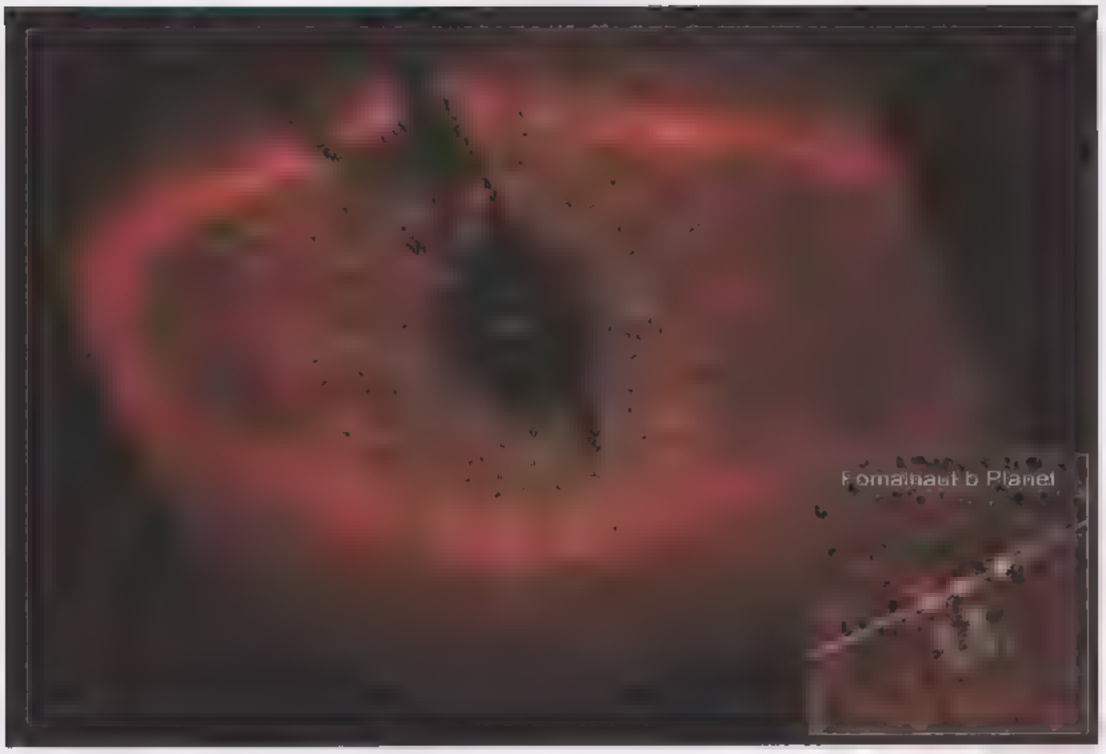


Figure 7.8 An image of the star Fomalhaut taken using a coronagraph on the Hubble Space Telescope. The black region near the centre is the occulted patch of sky centred on the position of the star (the white dot). The radial streaks are scattered starlight. The red dot at lower left is a background star. The small white box at lower right indicates the position of the proposed planet Fomalhaut b, which has carved a path along the inner edge of a vast, dusty debris ring encircling Fomalhaut. The identification as a planet was confirmed by its motion over 21 months, shown in the inset box, giving a mass of 2.1 M_J (i.e. 2.1 times the mass of Jupiter) and orbital distance of 116 AU. However, later non-detections in the infrared, where the 'planet' would be expected to be relatively brighter, have cast doubt on its reality. This example illustrates the extreme difficulty in obtaining exoplanet observations.

■ Equation 7.2 can still work, even if one object is not a planet. In this case, the left-hand side of Equation 7.2 would refer to one of the pair, while the right-hand side would refer to the other. What would happen if the two objects are exactly the same mass, as you might find in a binary star system?

The masses would be the same, so in order for Equation 7.2 to work, they must both be the same distance from the centre of mass. The stars would therefore stay on opposite sides of this point, while they each circle round it.

The mass of a star is much higher than the mass of a planet so the displacement of the star relative to the centre of mass and its orbital speed are both small compared with the distance and speed of the planet. However, it is still possible to detect the motion of a star due to its unseen planet.

The *astrometric method* involves measuring the tiny shift of a star's position relative to background stars, which are assumed to be in fixed positions. It is possible to determine the positions of the *centres* of stellar images to an accuracy of a small fraction of the size of the image itself. While this technique has proved successful in determining masses of binary stars, the tiny displacements in a star's position resulting from a planet (less than a milli-arcsecond for the example in Section 7.6.1 of a Jupiter-mass planet at

10 pc) mean that no planet has yet been discovered using this technique. Although advances in ground-based techniques are making such detections possible, the future for this technique lies with space missions. The space telescope GAIA, which is due for launch within a few years, is planned to detect motions as small as 20 micro-arcseconds.

In contrast, the *radial velocity method* has proved very successful. It uses the tiny Doppler shifts (Section 2.5.3) of lines in a stellar spectrum over time to detect the motion of the star about the centre of mass of the planetary system. As the star moves towards the observer in its orbit the spectral lines are shifted to shorter wavelengths and as it moves away, they are shifted to longer wavelengths (see Figure 7.9).

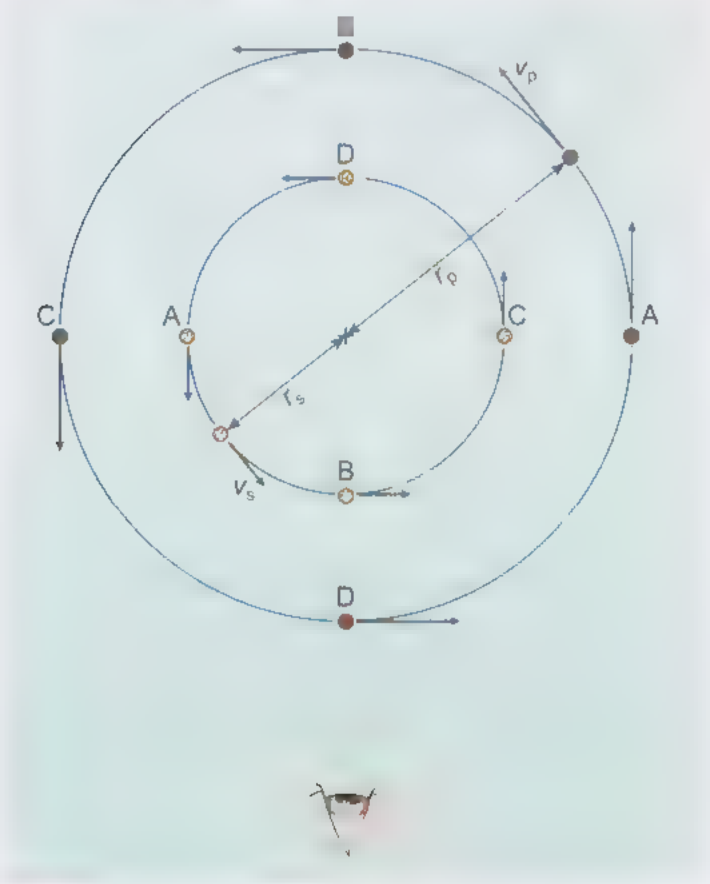


Figure 7.9 The orbits of a star (inner circle) and a planet (outer circle) around their centre of mass (marked 'x'). The astrometric method of planet detection involves measurement of the change in position of the star relative to background stars. The radial velocity method involves measurement of the orbital speed of the star v_s. The star will be moving towards the observer when at position A and away from the observer at position C.

Observations over an entire orbit result in an apparent oscillation of the lines due to the orbital speed, about an average wavelength that corresponds to the steady motion of the system through space. This is an identical technique to that used to derive masses of binary stars except that only one set of spectral lines is visible and the speeds and hence wavelength changes are much smaller. While a typical binary star may have an orbital speed of several tens of kilometres per second, resulting in a wavelength shift of around 1 Å at visible wavelengths, the orbital speeds of stars in extrasolar systems are a few tens of metres per second or less, resulting in wavelength changes of less than 0.001 Å. However, advances in instrumentation and observational techniques have made such observations possible, and this technique can be used for any stars that have enough narrow spectral lines (i.e. spectral types G, K and M) and are sufficiently bright to obtain high quality spectra. With current telescopes this limits the technique to stars within about 2000 pc. Although the details need not concern us here, the mass and orbital distance of the planet can be derived from the measured radial velocity and period if the star's mass is known. However, there is one further significant limitation; the measured maximum speed is only equal to the orbital speed if the orbits are aligned with the observer's line of sight (see Figure 5.4). As the inclination of the orbit plane to the line of sight may have any value, and cannot be determined from the observations, the measured speed will be lower than the true value and the derived planetary mass will be too low. The apparent planetary mass derived using this method is therefore a lower limit to the true mass.

Stellar systems are likely to have several planets so the motion of the star will be more complex. Precise observations can reveal the superimposed influences of several planets if they are sufficiently massive or close to the star.

7.6.3 Detection using changes in brightness

A planet may reveal its presence through changes in the observed brightness of a star.

The *transit method* involves the detection of the tiny change in brightness (less than 1%) as a planet passes in front of a star (Figure 7.10). The technique is limited to planets that have orbital planes very close to the line of sight. Although this may represent a small fraction of stars with planets, it is possible to observe many thousands of stars at a time with small telescopes. If the same area of sky is repeatedly observed, any transit can be observed if the change in brightness of a star is large enough. The orbital period is obtained from repeated transit detections. This method allows the sizes of planets to be determined from the amount of starlight that is eclipsed. The change in brightness for an Earth-sized planet around a Sun-like star is only 0.01%, which is beyond the capability of terrestrial telescopes due primarily to the effects of the atmosphere. Space missions provide the opportunity to search for transits of ever smaller planets. CoRoT (launched in 2006) and Kepler (launched in 2009) are currently flying and further missions employing the transit technique are planned. Once detected, radial velocity observations are required to determine the mass of the planet.

Einstein's theory of gravity, known as general relativity, predicts that massive objects distort space and so light can be 'bent' when it travels close to massive objects. Gravitational lensing by a nearby cluster of galaxies can lead to multiple or distorted images of distant galaxies (see Figure 3.7). On a smaller scale, individual stars or planets can cause tiny changes in the direction of light from a background star. This forms the basis of the *microlensing method* for planet detection. Stars are in constant motion in their orbits about the centre of our galaxy so if a star passes directly in front of another from our viewpoint, the nearby star can cause gravitational lensing of the light from the distant star, causing it to increase in brightness (Figure 7 11). If the lensing star has a planet in the correct geometry, then the

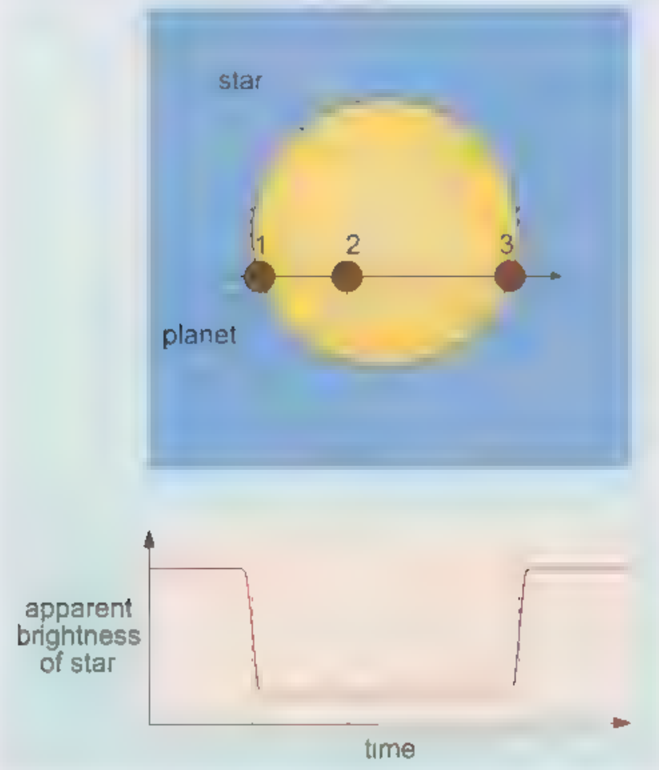


Figure 7.10 A planetary transit. As the planet passes in front of the star in its orbit as seen from Earth (upper panel) it causes an apparent dip in brightness of the star (position 1), illustrated in the light curve (lower panel). This dimming continues (position 2) until the planet passes out of the line of sight (position 3).

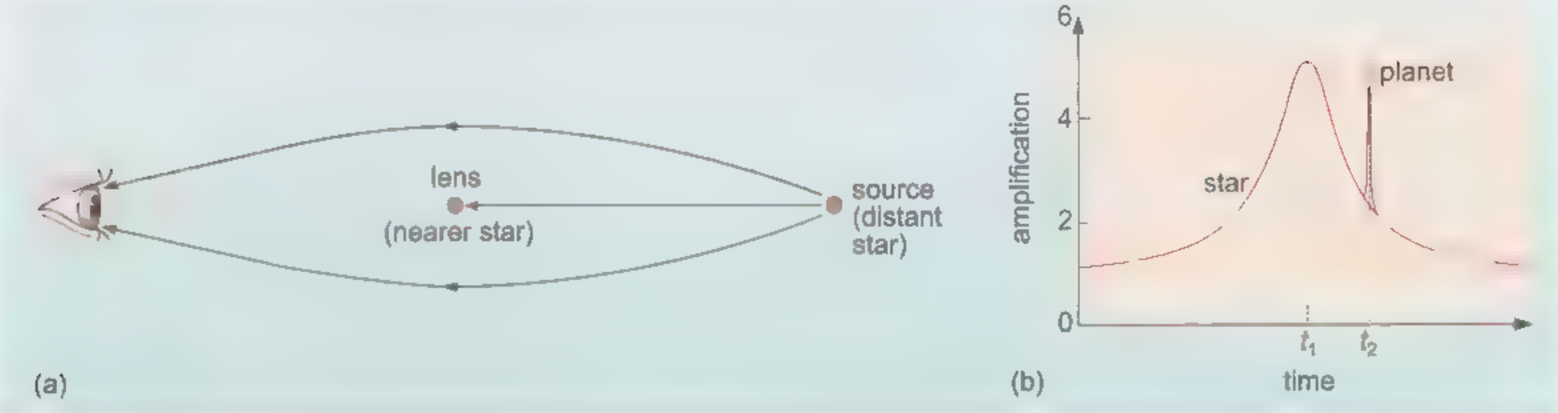


Figure 7.11 Gravitational microlensing occurs when the movement of stars causes an alignment with the line of sight to an observer (a) The path of light from the distant star is bent by the foreground star at time t_1, producing a lensing effect (b) The light curve of the microlensing event The combined light from the two stars is enhanced as the alignment occurs A planet orbiting the nearer star may also produce a second increase in brightness (shown by the black spike in the curve at time t_2).

planet's gravity can produce an additional enhancement in brightness from which the mass and separation of the planet can be deduced. Because the stars, as viewed from the Earth, appear too close for them to be distinguished, it is impossible to predict such lensing events, so as with the transit technique, many stars must be observed to search for lensing events. One disadvantage of this method is that transits do not repeat so the planetary orbits cannot be determined. Only a small number of planets have been detected to date using this technique but they include the lowest mass planets found using terrestrial telescopes.

7.6.4 Detection using timing methods

The first confirmed exoplanet discovery was made using a method not originally intended for planet detection. Pulsars are neutron stars that produce extremely regular pulses of radio waves as they rotate (see Section 6.6.2). If they have a planet, their motion about the centre of mass will cause variations in the timing of the pulses from which the planet's orbit can be determined. Although the *pulsar timing method* can potentially be used to detect terrestrial-mass planets, the extreme environments around pulsars and during their formation make them unsuitable habitats for life.

The presence of additional planets, in a system in which a planet has been detected using the transit method, can be determined by precise timings of the transits. Variations in the interval between transits (*transit timing variation method*) can be caused by the gravitational influence of other planets, possibly as small as terrestrial planets. Variations in the duration of transits could be used to infer the presence of a terrestrial-sized moon (an exomoon) about a Jupiter-sized exoplanet (*transit duration variation method*).

7.7 Properties of extrasolar planetary systems

The official definition of a planet by the IAU (Section 7.2.1) applies to the Solar System. What is important for extrasolar planets is the maximum mass a body can have, above which some thermonuclear reactions can occur and it is no longer classified as a planet. Above about 13 M_J (13 Jupiter masses) thermonuclear fusion of deuterium can occur and the object would be known as a **brown dwarf**. However, this mass refers to a body of stellar-like composition, i.e. predominantly hydrogen and helium, which would be expected if it formed directly from interstellar material. Objects that formed around protostars may have large rocky cores and hence may have significantly larger masses without undergoing deuterium fusion. Masses derived using some detection techniques may be uncertain or be lower limits Candidate exoplanets with apparent masses less than 13 M_J may therefore turn out to be brown dwarfs and some with masses larger than 13 M_J may turn out to be planets. For these reasons, exoplanet catalogues often list objects with masses significantly higher than 13 M_J.

Figure 7.12 summarises the masses and orbital distances of exoplanets discovered to date. This diagram tells us as much about the limitations of the different methods as it does about the properties of exoplanets The transit method favours large planets (which produce deeper dips in brightness) that are close to their parent star (the short orbital period allows repeated transits in a short timescale to help confirm the discovery). The transit method can only detect planets that have orbits close to the line of sight, which also favours planets close to the star. Only a few percent of planetary systems will produce transits. The radial velocity method also favours massive planets orbiting close to the star, which produce the largest oscillation in the star's radial velocity. The radial velocity method is restricted to stars that are bright enough to obtain high quality spectra and have narrow spectral lines If the plane of the orbit is nearly perpendicular to the line of sight, the radial velocity will be too small to measure. Radial velocity searches have also tended to focus on Sun-like stars (single stars of spectral type F, G or K). Detection with direct imaging favours large but distant planets.

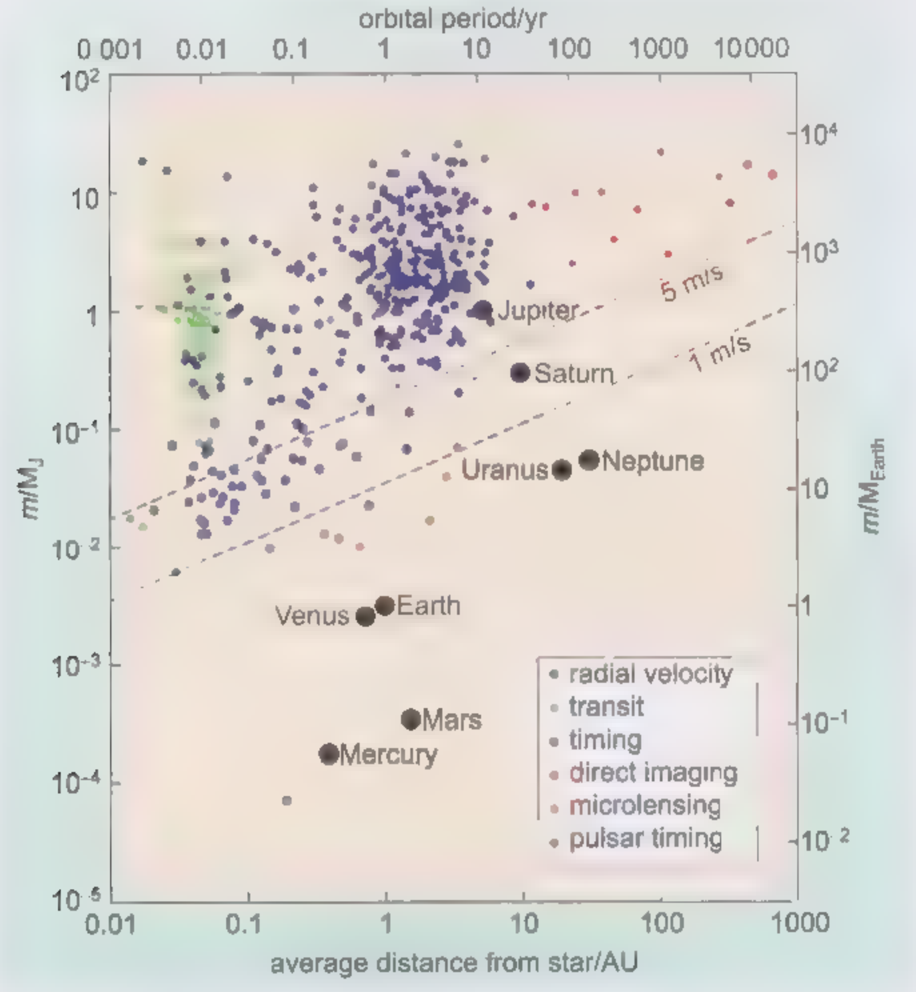


Figure 7.12 Properties of exoplanets discovered using different techniques The lower axis gives the average orbital distance and the vertical axes give the mass (left-hand axis is in units of Jupiter's mass and the right-hand axis in units of Earth's mass). The top axis indicates the orbital period (calculated assuming the star is a solar type star). The sloping dashed lines indicate the variation in radial velocity (of a $1\,M_\odot$ star) from a planet at that location in the diagram.

The first exoplanet discovered orbiting a main-sequence star, 51 Pegasi b in 1995, was a big surprise. It had a mass close to that of Jupiter, but orbited its Sun-like star in only 4.2 days. This gave an orbital distance much less than that of Mercury in the Solar System. Further discoveries revealed Jupiter-mass planets with even shorter periods. Although the discovery techniques favoured these properties, such massive planets so close to their host stars were not expected to exist. It is generally accepted that giant planets form around massive ice cores beyond the snow line in the nebula around the newly forming host star. Such icy cores could not form so close to a star, but the planets clearly exist. The conclusion was therefore that these **hot Jupiters** formed further from their host stars and have subsequently migrated to their current positions. Some are so close to their host star that the high temperatures will result in steady loss of their atmospheres.

The orbits of many of the exoplanets appear to be eccentric, have their orbital planes tilted compared with the star's rotation, or even have retrograde orbits. Although some of these results may be caused by confusion with multiple planet detections, they do not fit with current ideas of planet formation which produce near circular, in-plane orbits as found in the Solar System.

A planet discovered with the transit method allows direct determination of its radius. If the radial velocity is then measured, the planet's mass can be determined because the orbital plane is known to be close to the line of sight With the size and mass known, the average density can be derived. Most densities measured so far lie in the range 300–2000 kg m^{-3}, consistent with gas giant planets, although the predominance of very low values is still puzzling astronomers It is difficult to estimate the proportion of stars that have planets because of the biases in the different types of discovery technique. However, if all these and other known effects are taken into account a few facts emerge:

- Most exoplanets orbit Sun-like stars even when the observational biases are taken into account. Lower-mass stars are either less likely to have planets or their planets are typically of lower mass and hence harder to find. Stars of very high mass are much rarer and the effects of their intense radiation may inhibit planet formation.
- A few percent of Sun-like stars have Jupiter-mass planets in close orbits and perhaps 20% have at least one giant planet.
- The number of terrestrial-planet discoveries is currently too small to estimate the numbers present, but if the Solar System is representative, then they may be very common. The Kepler mission, launched in 2009 is expected to provide the answer to this question. Kepler results in January 2012 already suggest that our galaxy may contain more planets than stars!
- Stars of higher metallicity (meaning, richer in elements that are heavier than hydrogen and helium) are more likely to have planets. They formed more recently from the interstellar medium that is enriched in heavier elements required to form planets.

A large fraction of detected exoplanets are hot Jupiters, because these are the most easily detected It is expected that there are many systems more like the Solar System, with terrestrial-mass planets orbiting at distances where conditions are conducive to life. The detection of such planets is one of the major goals of astronomy today. Once detected, it is a further challenge to identify whether life may be present. The next chapter will investigate where life may exist elsewhere in the Universe and how one might deduce its presence.

End-of-chapter questions

Question 7.1

Figure 7.3 illustrates why seasons on the Earth occur. Use the information in Table 7.1 to (a) compare the properties of the seasons on Uranus with those on the Earth, and (b) explain why seasons on Mercury will have very different characteristics from those on the Earth.

Question 7.2

What are the main characteristics of the terrestrial planets that distinguish them from the giant planets?

Question 7.3

Which features distinguish an impact crater from a volcanic crater?

Question 7.4

If we observed the Solar System from a distance of 10 pc, how much fainter would Jupiter appear than it does from Earth? (See Section 1.5 if you need to remind yourself of the inverse square law.)

Question 7.5

In Table 7.4 each column is a property of an exoplanet or its host star and each row is a different method of exoplanet detection. Complete each empty location in the table to indicate which values of a property are more easily detected by the method (e.g. 'high' or 'low', 'large' or 'small'. If there is no preference, or the property is not a defining factor then enter '–'. One row is completed as an example. Although stars detected by the radial velocity method may be generally within 2000 pc, it is the apparent brightness of the star that determines its detectability.

Table 7.4 Preferred properties for exoplanet detection methods.

	Planet's mass	Planet's size	Planet's orbital distance	Distance from Earth	Star's mass	Star's apparent brightness
Direct imaging						
Astrometric method						
Radial velocity method	high	–	small	–	low	high
Transit method						

Now go to the module website and do the remaining activities associated with Chapter 7.

Activity

Chapter 8 Life in the Universe

8.1 Introduction

Some scientific questions are of such immediate and widespread interest that they stir all our imaginations. One of these questions is the subject of this chapter – is there life beyond the Earth? This question has surely been pondered ever since people in antiquity realised that there might be celestial bodies of rock and water beyond the Earth, but it is only in the last 50 years or so that significant scientific progress has been made towards answering it Today, the question is still unanswered, but it is the focus of intense scientific activity. This is a fast-moving area with space missions planned to continue the exploration of likely habitats in the Solar System and the search for Earth-like planets beyond it.

In order to search for life elsewhere in the Universe it would be useful to have a definition of life so that we could recognise it. This is by no means as simple as it sounds, even for life 'as we know it'. If we observe our surroundings we would all expect that we could judge what is alive (plants, animals, insects) and what is not (e.g. rocks, water, buildings, machinery). But, as the famous astrobiologist Carl Sagan said, " '*I'll know it when I see it*' is not good enough". There have been many attempts to define precisely what is alive. Any successful definition would have to be flexible enough to encompass everything from the simplest bacteria and single-cell plant life to complex plants and animals. Such definitions have generally been based on combinations of attributes such as the ability to grow (involving taking in energy and nutrients and expelling waste products), to respond to external stimuli, to reproduce, adapt and evolve. More advanced life forms may be mobile, communicate and interact deliberately to change their environment. But there are difficulties with all of these definitions. A mule, the offspring of a horse and a donkey, cannot reproduce, but it is alive. Viruses evolve and replicate but require another organism to provide the nutrients for growth. Machines have been designed to build new machines, robotic devices respond to external stimuli and change their environment and 'intelligent' software learns and evolves from experience, although none of these is alive.

Even with a perfect definition of life, it may also be a challenge to identify it elsewhere in the Universe, particularly if its location is beyond the reach of space exploration The first step is to determine the requirements for life to exist, which makes it possible to ask the following questions:

- Are there places elsewhere in the Universe that life can develop and survive?
- What are the most likely places?
- Can extraterrestrial life be discovered or observed?
- Can we (human beings) find evidence for life?
- What about intelligent life?

The Earth is the one place in the whole Universe where we *know* there is life. The search for life elsewhere must be guided by the key features of life on

Earth. Section 8.2 considers the requirements for survival of life as we know it and how these conditions are satisfied on the Earth Section 8.3 investigates which other bodies in the Solar System might have been suitable for the emergence of life, or might support life today. Section 8.4 takes a look beyond the Solar System. We (humanity) can use our knowledge of stars, extrasolar planetary systems and life in the Solar System to deduce where life may survive elsewhere in the Universe.

A few people believe that they already *know* that there is life beyond the Earth; they claim to have been visited by aliens. Although it is easy to laugh at the notion of alien visitation, it is in fact perfectly serious to ask whether it has actually happened, recently or in the more distant past. The perfectly serious answer is that there is no scientific evidence for it. The claims lack objective facts Certainly, there are unexplained happenings, but the important word is 'unexplained', and a belief that alien visitation is the explanation is just that, a *belief*. Although life may be common in the Universe, *intelligent* life may be very rare and the difficulty of travel over the vast distances between stars makes the possibility of an alien visit very small indeed. The possibility of intelligent life elsewhere and how we might communicate with it is considered further in Section 8.5.

8.2 Requirements for life

Life on Earth today displays bewildering variety, yet all organisms in a fundamental sense are rather similar. The most familiar organisms are large ones – from elephants, blue whales and trees, down to spiders and fleas Yet even at the small end of this size range an organism consists of a large number of what are called cells – about 10^6 in the case of a flea. Other organisms consist of just a single cell – bacteria are a widespread example There is a huge variety of bacteria. Although bacteria are commonly associated with disease, most of them are harmless and many have vital roles in the health of the biosphere, such as recycling important nutrients for life. They account for a large proportion of all single-cell organisms, and such organisms may have accounted for *all* life on Earth from its origin nearly 3900 million years ago until as recently as about 600 million years ago. Multicellular organisms would therefore be a comparatively recent development. A few single-cell organisms are shown in Figure 8.1, where each cell is roughly a few hundredths of a millimetre across A cell consists of a membrane that encloses liquid water that contains some other substances essential for the cell to grow and reproduce. It is these substances in the cell that define the similarity between all organisms, substances that are common to all life on Earth, and that are essential for all life on Earth.

- If we were trying to 'build' a living organism, what would be required?

It would require materials to build the structure (cells) of the organism, a transport mechanism to move these materials to the required locations and energy to perform the construction. In addition, the environment in which the organism is to live must provide stable conditions so that the organism can survive and grow.

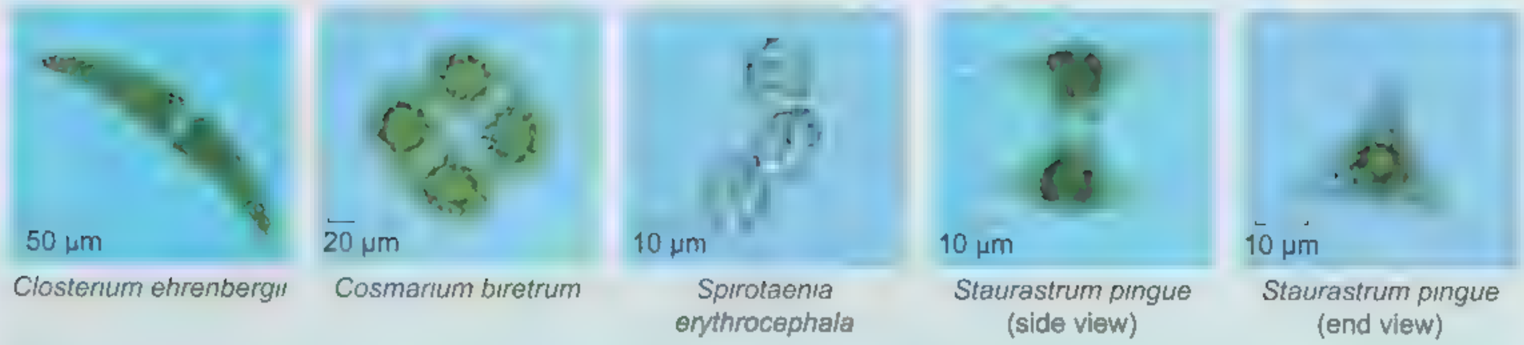

50 µm
Closterium ehrenbergii
20 µm
Cosmarium biretrum
10 µm
Spirotaenia erythrocephala
10 µm
Staurastrum pingue (side view)
10 µm
Staurastrum pingue (end view)

Figure 8.1 A variety of single-cell organisms. In the case in the centre there is a colony of three single cells.

8.2.1 Molecules for construction

Constructing the cells of a living organism needs a wide variety of building materials molecules to provide all the complex components that allow it to survive and grow. If you looked into a cell, at the fundamental chemicals of life, you would find that all life on Earth is based on huge molecules that are complex compounds of the chemical element carbon (C). Carbon can form chemical bonds with many other atoms, allowing a great deal of chemical diversity. The elements hydrogen (H), oxygen (O), nitrogen (N) are the most common atoms found in carbon-based molecules (*organic* molecules). It comes as no surprise to chemists that carbon is the basis of life. No other chemical element comes anywhere near carbon in its ability to form the large and complex compounds that are necessary for life.

Proteins are complex organic molecules that provide structure, for example in human hair or the walls of individual cells. There are around 100 000 types, each made of 50 to 1000 smaller components called amino acids, that provide the variety of structures required. Amino acids can be regarded as the building blocks of life. Only 20 amino acids are found in proteins of living systems but they are combined in an enormous variety of ways. The most famous large organic molecule is DNA (deoxyribonucleic acid, see Figure 8.2), which stores the genetic code, the instructions for building proteins, for all types of life on Earth.

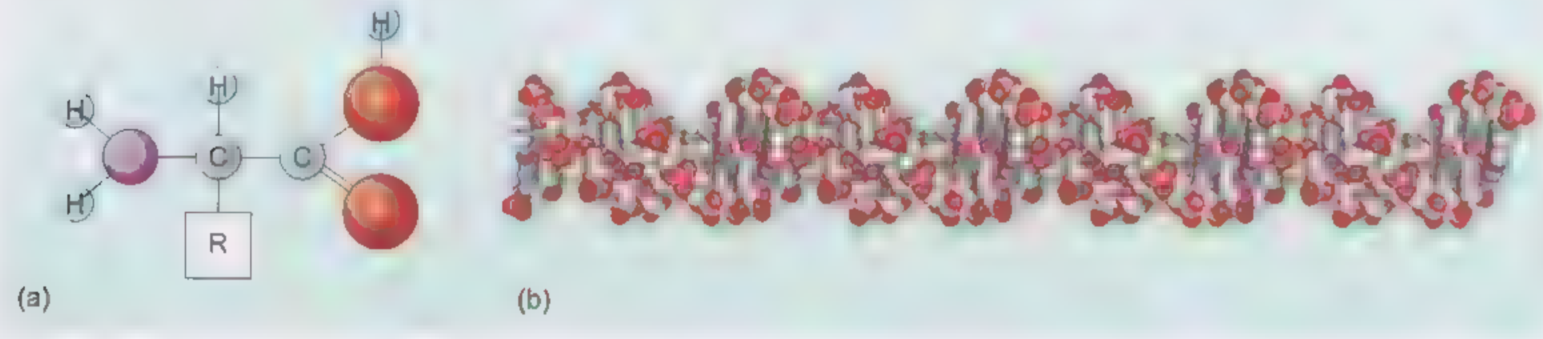


Figure 8.2 Examples of organic molecules required for life on Earth: (a) A class of amino acid where ‘R’ may be a single hydrogen atom or a more complex organic component. (b) Part of a DNA molecule.

These and other types of giant molecules, called macromolecules, provide the structure of cells, are used to store and release energy, contain the information for reproducing new cells and aid in chemical reactions, which maintain the survival of the organism. However, without one very simple molecule, most of these functions would not be possible. That molecule is water.

8.2.2 Water

The majority of the mass of a typical cell is water. Many organic molecules dissolve easily in water so it provides a medium in which molecules mix so that chemical reactions can occur and the products can be transported where they are needed. Water appears to be an essential requirement for life. Cells can be thought of as bags of molecules, separated from the outside world by a wall or membrane that keeps water and organic compounds together.

In order for water to provide this vital function, it must be liquid. Although there are some single-cell organisms that can live in snow, they still rely on melted snow to grow. This restricts the range of conditions in which the cell can operate. If those conditions are such that water is in its solid form, ice, then the cell cannot function and if it boils to form water vapour it will destroy the cell. As you will see shortly, the need for liquid water is one of the greatest constraints on the environments in which we might find life.

8.2.3 Energy

Energy is required to power the processes that maintain life In humans, this energy comes from food. The energy is stored in certain types of organic molecules that are themselves produced by other plants or animals Originally, this stored energy came from a non-biological source.

- ■ What is the most abundant source of energy on the surface of the Earth?
- □ Sunlight provides over 1000 W of power for every square metre at the distance of the Earth. Even with a significant fraction reflected from clouds it still dominates the energy budget at the Earth's surface.

Plants convert carbon dioxide, water and other nutrients into more complex organic molecules by photosynthesis, effectively converting the energy of sunlight into chemical energy stored in these molecules. Some organisms are found in habitats where there is no available sunlight, such as deep within rocks or in the depths of the Earth's oceans, so they need a different energy source.

Figure 8.3 shows one such source of energy, the release of heat stored in the Earth's interior. These vents occur in regions where ocean water has seeped into fractures in the Earth's crust, is heated by magma at a temperature of over 1000 °C and is forced back up to the sea floor. Despite the complete absence of sunlight at these ocean depths, life flourishes, using the hydrothermal energy and nutrients from the dissolved minerals carried up by the extremely hot water.

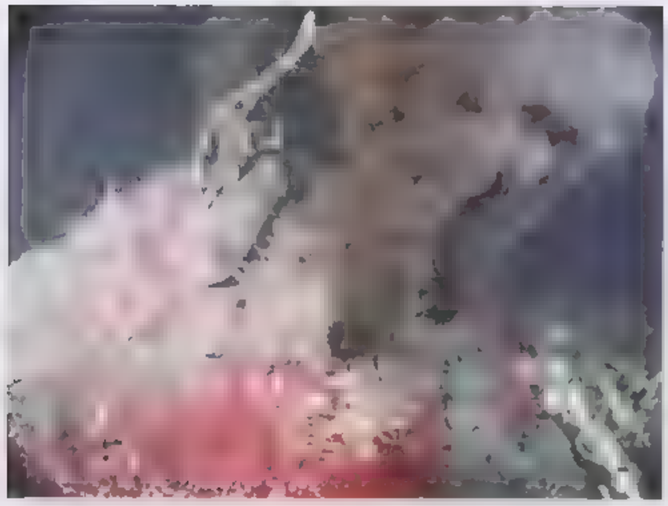


Figure 8.3 A hydrothermal vent in the Pacific Ocean.

8.2.4 Environment

The preceding sections have identified the three main requirements for life of the kind found on Earth: (i) carbon to form the complex organic molecules that perform the many functions required in a living organism, (ii) liquid water to allow the reactions of these molecules to take place and (iii) energy to power the process. However, there is one further requirement for us to find life· the environment in which these components are present must remain stable for long enough for life to develop and survive.

Carbon is a fairly common and widespread element but complex organic molecules can exist only under certain temperature conditions and this is one of the most restrictive properties for the survival of these molecules.

- Make an 'educated guess' about what might happen to these carbon molecules at sufficiently high temperatures.

  They will break up.

Specifically, the molecules that are the basis of life break up at temperatures above about 150 °C, so we must confine our attention to places that are colder than this. We have already seen that for organisms to grow, we require *liquid* water. We must therefore identify places where water exists, and where the temperature and pressure is such that the water is liquid.

- At sea-level on the Earth, over what range of temperatures is water liquid?

  Water is liquid between 0 °C and 100 °C at sea-level. Thus, if the temperature is lowered to 0 °C water freezes, and if it is raised to 100 °C it turns into vapour (gas) very rapidly, i.e. it boils.

These boiling and freezing temperatures depend on atmospheric pressure. At sea-level on the Earth the average pressure is about 1 bar, or 1000 millibars (the millibar is a unit of pressure that you might have seen on weather maps). At 1000 millibars the boiling temperature of water is 100 °C. On a mountain only 1000 metres high, the average pressure is reduced to about 900 millibars and the boiling temperature is consequently reduced to 97 °C. The boiling temperature is raised above 100 °C in a pressurised container with a pressure of greater than 1 bar. The freezing temperature of water is much less sensitive to pressure   for practical purposes we can regard the freezing temperature as 0 °C.

As the pressure is lowered below 900 millibars, the boiling temperature of water continues to fall, and approaches 0 °C at a pressure of 6.1 millibars. Consequently, below 6.1 millibars, water cannot exist as a liquid, but only as a solid or a gas. For practical purposes, we can assume that if the conditions allow liquid water to exist, then complex carbon compounds can exist too.

However, heating is not the only process that can destroy organic molecules. They may be damaged or destroyed by chemical reactions in very acidic or alkaline environments, or in the presence of highly reactive chemical compounds. Even in very harsh environments on Earth, organisms have evolved ways in which they can protect themselves from such effects (see Section 8.2.5). High-energy photons of light and energetic charged particles can also break up organic molecules.

■ The Sun produces high-energy X-ray and ultraviolet photons, and charged particles, so how can large organic molecules survive on the Earth?

The Earth's upper atmosphere absorbs the high-energy photons (see Figure 2.4). The Earth's magnetic field traps or deflects the charged particles. During solar flares or periods of enhanced solar activity, when the particles do reach the Earth, they interact with the upper atmosphere, causing aurorae.

The majority of solar radiation is emitted in visible light but its intensity at ultraviolet wavelengths is sufficient to damage organic molecules. We are protected by a high-altitude layer of ozone (O_3), which is formed with the help of the break-up of oxygen molecules (O_2) by energetic photons, and absorbs radiation of wavelengths less than 320 nm. Without the protection of the Earth's atmosphere and magnetic field, most organisms could not survive.

Even if a planetary environment is conducive to life, the favourable conditions need to be maintained long enough for life to form and evolve. Although it is extremely difficult to determine the date of the first life on Earth, it took more than 3 billion years for animal life to evolve and roughly a further 600 million years for intelligent life. From the one example we know, the environment must remain relatively stable for millions or billions of years for advanced life to develop. Nevertheless, once established it seems that life is very tenacious. During the Earth's history there have been a number of mass extinctions, when a significant fraction of all living species on the Earth disappeared. There is still debate over the causes of these extinctions, but the current consensus is that impacts of km-sized asteroids and the consequent climate changes from dust in the atmosphere are responsible for most of these extinctions and perhaps all of them. Atmospheric changes as a result of enhanced volcanic activity may have also played a role. Despite these and other climate changes, life has prevailed – and its evolution may have even been enhanced by the stimulus of environmental change. Other rare astronomical events could also potentially cause elimination of life. For example any planet orbiting a star close to the site of a supernova would be sterilised by the intense burst of radiation. The survival of life on astronomical timescales may be at the whim of a cosmic lottery!

8.2.5 Life in extreme environments

The search for life on other planets is necessarily constrained by what is known about life on Earth, since it is the only example we have. To help guide the search for life, scientists visit environments on Earth that are so unusual (in some physical or chemical extreme) that we may reasonably guess they resemble extraterrestrial environments. Organisms that can survive in these extreme environments are called, not surprisingly, **extremophiles**.

The cold deserts of Antarctica offer insights into how life can survive in extremely cold conditions. As the average temperature on Mars is about −55 °C, this might be a good place to understand how organisms produce chemicals that help them survive freezing and thawing. In the hot deserts of Chile, the extreme desiccation helps scientists understand how life might survive on the surface of planets where liquid water is very scarce. As liquid water is extremely scarce near the surface of Mars, if there is any life there, it would certainly have to be able to cope with desiccation.

When scientists go to these environments they look for protected habitats places where life might have some opportunity to survive the extreme conditions and flourish in environments that are otherwise hostile to life. An example of such habitats is a 'cryptoendolithic' habitat literally a habitat hidden within rocks Some micro-organisms can invade the cracks and pore spaces within the rocks and live within the material. You can see such an example in Figure 8.4. This is an example of a cryptoendolithic community of cyanobacteria living in a rock in the Arctic. The micro-organisms are photosynthetic, which means they need light for their energy needs As you can see, they have to grow at a depth in the rock where the light levels are enough for photosynthesis. Too low and they don't get enough light. By growing inside the rock, they escape the extremes of the surface of the rock, which include exposure to solar ultraviolet radiation and desiccating winds The result is a distinct band within the rock where the organisms can grow.

These habitats in rocks are found in many extreme deserts of the world, both hot and cold deserts. In addition to living in rocks, micro-organisms can also live under rocks, where they get the same protection from environmental extremes. As photosynthetic micro-organisms need light, to live under a rock requires that the rock itself is translucent – allowing sufficient light to penetrate through the rock to reach the micro-organisms beneath Quartz is a common rock that is colonised on its underside in extreme deserts. Organisms that live on the underside of rocks are called 'hypoliths'. Of course, it's not known yet whether such habitats are colonised on other planets such as Mars, but studies of these habitats can reveal how life survives in extremes on Earth and whether there is any likelihood of life surviving the extremes that are found elsewhere.

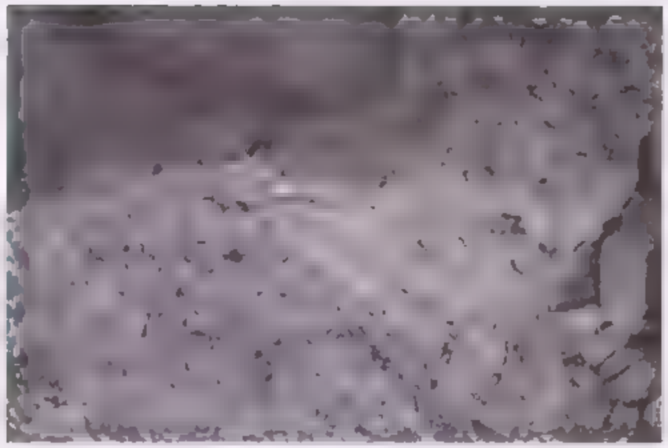


Figure 8.4 A cryptoendolithic community of cyanobacteria living in a rock from the Arctic. Note the scale bar in the bottom left corner.

The search for so-called 'analogue' environments is revealing many facts about life on Earth. Several decades ago, it would have been inconceivable that life existed in the deep oceans or even in the clouds. However, studies of these environments and many others have revealed the remarkable versatility and tenacity of micro-organisms, including their ability to harness energy supplies in the form of light and chemical compounds and their ability to survive and grow in extremes of temperature, acidity, salinity and radiation, among other things. In turn, these studies are aiding scientists in the search for life on other planets.

8.5 Where might we find life in the Solar System?

Our examination of the requirements for life indicates that we need to search for environments that provide conditions conducive to the existence and survival of large organic molecules and the presence of liquid water.

Activity 8.1 Which planets in the Solar System could support life?

The estimated time for this activity is 20 minutes

On the basis of the information given in Tables 7.1 and 7.2 and Section 7.3, make a list of each of the planets (not including the Earth) and dwarf planets and state, with reasons, whether each one is likely to be a potential extraterrestrial habitat. In some cases you might have to decide that there is insufficient information to form any sort of judgement.

From the comments on Activity 8.1 it is clear that, among the planets and dwarf planets, none (except the Earth) is very promising as a potential habitat at the present time. This is borne out by further data on the planets. Only one planet has any realistic chance of being a potential habitat, and that is Mars.

However, there are further candidates among the large satellites of the gas giants as you will see.

8.3.1 Mars

Mars is the next planet out from the Sun after the Earth. It is a rocky body, with a diameter about half that of the Earth, and rotation period and obliquity almost the same as Earth's. When viewed through a telescope it reveals white polar caps that grow and shrink with the seasons and seasonal colour changes that could be interpreted as growth of vegetation. Speculation about the possibility of intelligent life on Mars reached a peak in the years around 1900 when a number of respected astronomers drew maps, from their visual observations, of a planet criss-crossed by straight linear features. These channels or 'canals' turned out to be illusory (Figure 8.5) and the cause of the seasonal changes was confirmed by space missions to be the result of seasonal global dust storms masking the dark features wrongly interpreted as vegetation.

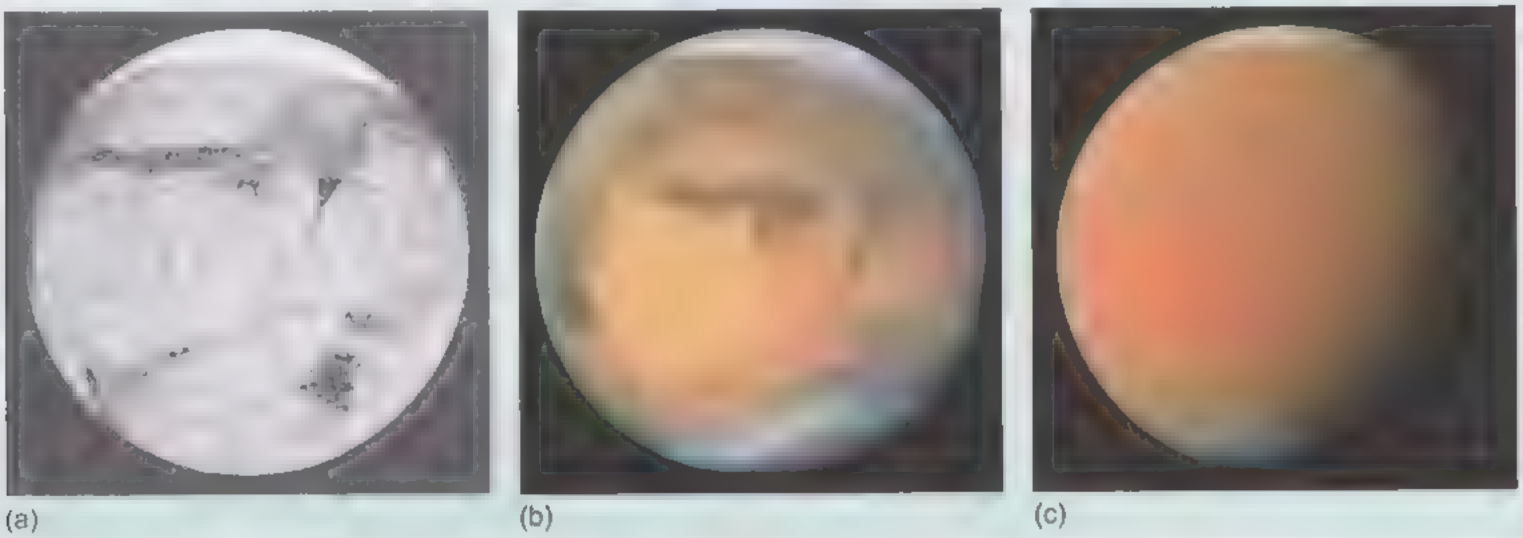


Figure 8.5 (a) The supposed channels, or 'canals', on the surface of Mars, drawn in 1905 by the American astronomer Percival Lowell, which led to speculation about intelligent life on Mars. (b), (c) Modern telescopic images of Mars at a similar geometry showing the effect of a seasonal global dust storm.

Mars has a thin atmosphere of carbon dioxide, and surface temperatures that can reach 20 °C on a very warm day. The average surface atmospheric pressure on Mars is close to the 6.1 millibar minimum for liquid water. Therefore, only in the very deepest chasms on Mars, where the pressure would be slightly higher, and even there only on the warmest days, could liquid water persist at the surface. The polar caps (see Figure 8.5), which are visible even in small telescopes from the Earth, contain both water ice and carbon dioxide ice.

Water-carved features were first seen in images of Mars acquired by the orbiting spacecraft Mariner 9 as long ago as 1971, providing firm evidence of flowing water early in Mars' history (2 to 4 billion years ago: Figure 8.6). Subsequent Mars missions have revealed abundant evidence for the flow of water much more recently. Minerals that indicate surface rocks were once soaked in liquid water have been detected from orbit and by landers on the surface of Mars. Gullies, believed to have been formed by liquid water flows less than 1 million years ago, have been found on the step walls of many craters and in some cases this flow has happened in the last few years (Figure 8.7). Unlike similar features on the Earth, these could not have been produced by rainfall but by water released from subsurface deposits, probably as the pressure of liquid water rose and burst through a permafrost layer. Although liquid water cannot survive at the surface, water ice has been found in abundance just below the surface at high latitudes and, in some cases on the surface itself (Figure 8.8). Even at the equator there is evidence of ice lying just below the surface (Figure 8.9).

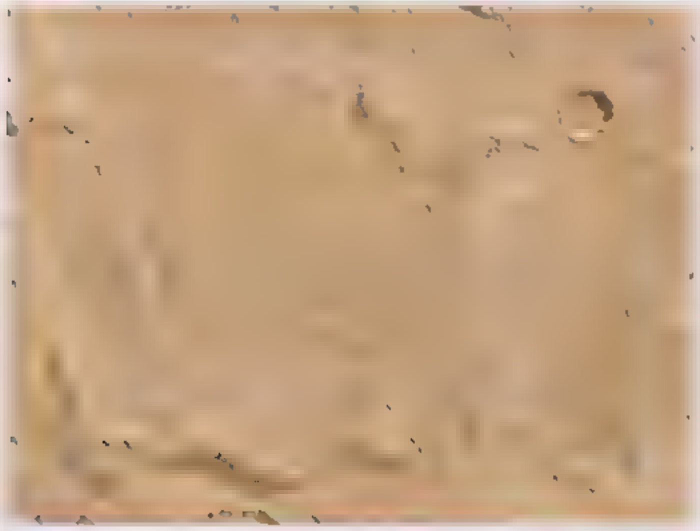


Figure 8.6 Ancient river valleys and impact craters on Mars.

Despite the presence of water ice at or near the surface and recent water flow, there is certainly no evidence for life at the Martian surface. No tracts of vegetation have been seen from space, and the spacecraft that have landed have not seen life-forms stalking the landscape. More significantly, the landers have seen a total absence of the processes and chemical products of current life in the Martian sands. Although very little of the Martian surface has been explored in this way, it seems unlikely that there is any life at the surface today.

- ■ The ozone layer and magnetic field are two characteristics of the Earth that help protect living organisms on its surface. How does Mars compare?

- □ From the information in Table 7.2 we can see that Mars' atmosphere does not contain abundant oxygen and therefore cannot form an ozone layer to protect the surface from high doses of ultraviolet radiation. The very weak magnetic field, coupled with the thin atmosphere also means that energetic particles can reach the surface.

If there is any current life on Mars then it must be below the surface and is most likely to be deep under the surface, where the pressures are large and where heat from the planet's interior could keep the temperatures above 0 °C, and thus water could exist as a liquid.

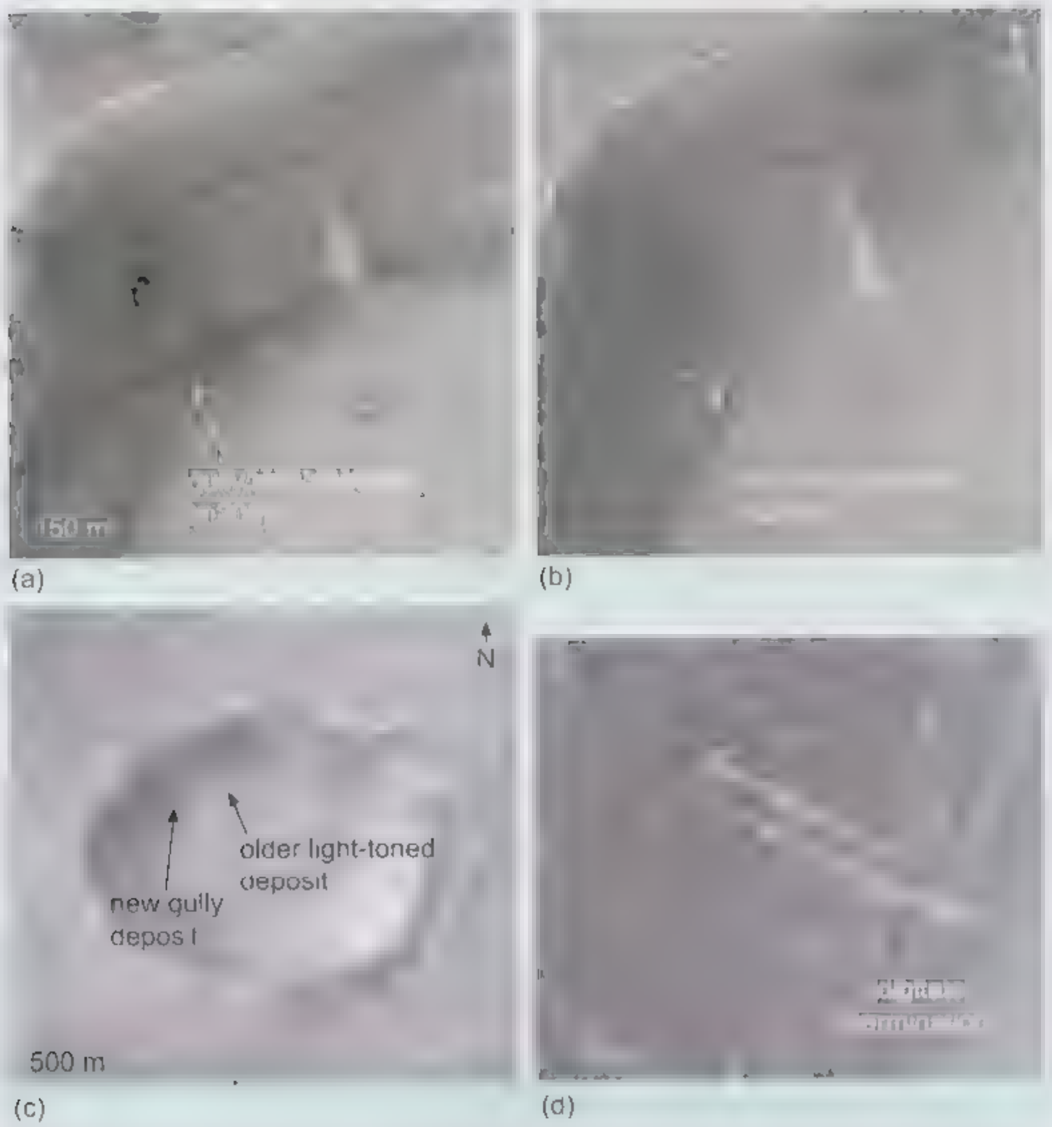


Figure 8.7 (a, b) Two images of the same area of a Martian crater (c) taken several years apart. A light deposit (enlarged in (d)) has appeared in the second image, indicating flow of material down a gully in the crater wall. These gullies, less than 1 million years old, were already believed to provide evidence for recent water flow. While these changes do not prove that water still flows on Mars, they provide the first very tantalising evidence that this may have occurred. Although the surface is extremely dry, the release of liquid water from ices beneath the surface may have caused this flow of material.

Thus there is a possibility of life existing deep in Mars, just as on Earth where there is life deep in the rocks, provided that liquid water can reach such places. However, most astrobiologists (scientists whose research interests are in extraterrestrial life) consider life deep in Mars to be a remote possibility. By contrast, they consider it much more likely to find evidence of life at the surface of Mars in the distant past, and that evidence for this will be found in the form of fossils or chemical evidence of past life. Although short-lived flows of water have occurred recently, it is the water that appears to have been common as a liquid on the surface in the distant past that may have provided conditions for life to become established. Clement conditions could have lasted long enough for life to evolve, although probably not long enough for it to evolve far beyond the single-cell stage.

How is it possible to determine the ages of planetary surfaces? In the next activity you can explore how ages are established for the various terrains of Mars.

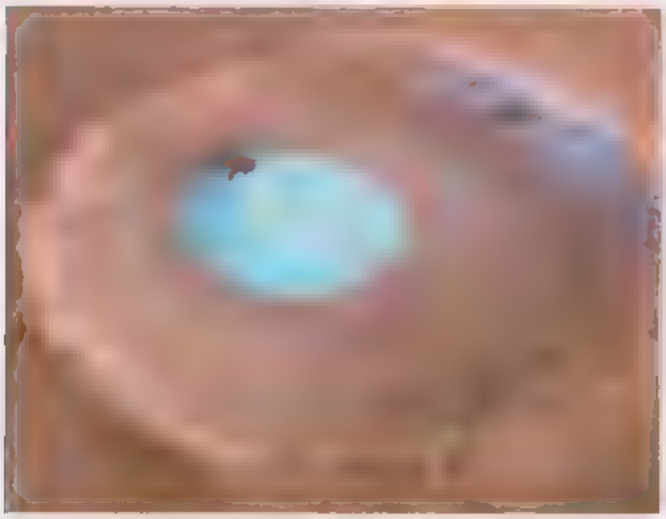


Figure 8.8 A lake of water ice in a crater near the north pole of Mars.

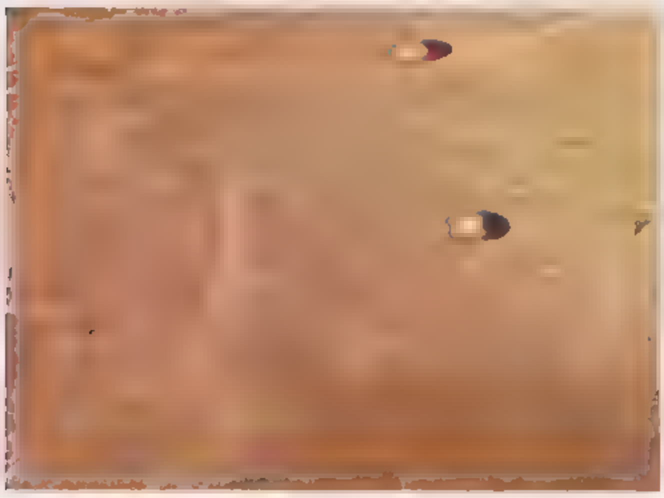


Figure 8.9 Structures resembling ice floes are apparent in this image, a few tens of kilometres across, of part of a vast plain near Mars' equator. Although ice cannot survive long at the surface near the equator, it may have been protected by volcanic dust after a catastrophic flood of the region within the last few million years.

Activity 8.2 Dating the Martian surface

The estimated time for this activity is 20 minutes.

Study Figures 8.6 and 8.9. How do the appearance and relative numbers of craters compare? From your comparison, explain how impact craters could be used to deduce that one terrain on the surface of Mars is older than another.

Outline a simple demonstration of your own invention that would illustrate the principle.

Mars hit the headlines in 1996 after a NASA press conference was convened to announce the publication of a scientific paper claiming that biological microfossils had been discovered in a meteorite from Mars known as ALH84001. This claim has been disputed, and the balance of opinion is that the tiny structures, and the accompanying chemical features of the meteorite, have a non-biological origin.

Another example of the difficulty of interpreting possible evidence for life lies in the recent detection of methane (CH_4) in the atmosphere of Mars (Figure 8.10). Methane is of interest to astrobiologists because on Earth most of the methane is released by organisms as they digest nutrients. However, other processes that are purely geological can also release the gas. Methane is destroyed relatively quickly in Mars' atmosphere, so whatever the subsurface source of the gas, its release implies recent biological or geological activity. The study of methane is one of the prime objectives of the Trace Gas Orbiter mission, currently planned for launch in 2016.

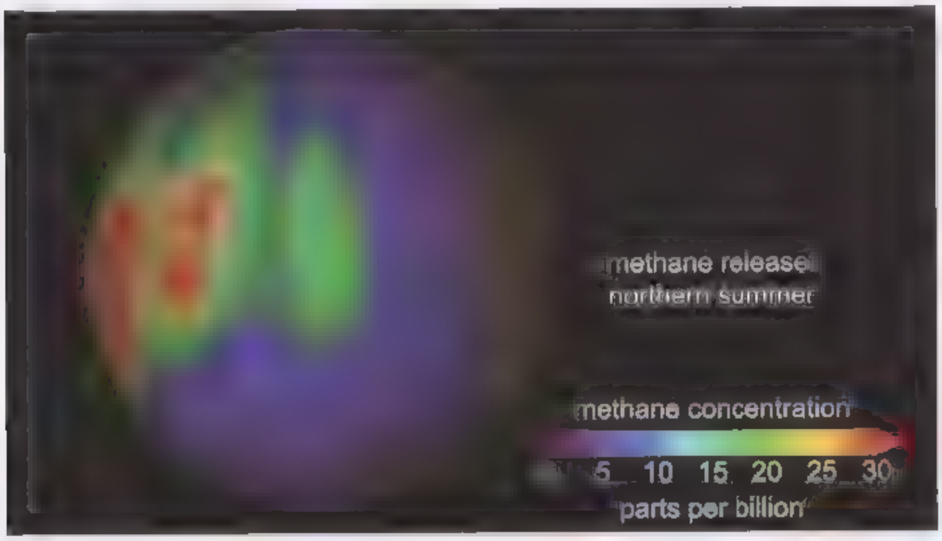


Figure 8.10 Methane release detected in the atmosphere of Mars. Red areas show the highest concentrations of methane and purple areas the lowest.

Mars has hardly been explored for evidence of *past* life, which was one of the objectives of the Beagle 2 lander (developed within the OU's Planetary and Space Sciences Research Institute). This lander was released from the European Space Agency's Mars Express spacecraft for a landing on Mars in December 2003 but unfortunately no signals were received after landing. More ambitious plans are underway for landers to search for evidence of life, present and past, and ultimately to return samples to Earth.

8.3.2 Europa

Europa is one of the four large satellites of Jupiter, each of which is about the same size as Mercury, so they would be regarded as planets if they were in their own orbits around the Sun rather than in orbit around Jupiter. However, the fact that these satellites *are* in orbit around Jupiter has led to the possibility of life being found there.

Why has this been proposed? All the gas-giant satellites reside at large distances from the Sun where surface temperatures are expected to be below −150 °C and, with the exception of Titan, they have no appreciable atmospheres. They appear very unlikely habitats for life. Figure 8.11 shows an image of Europa.

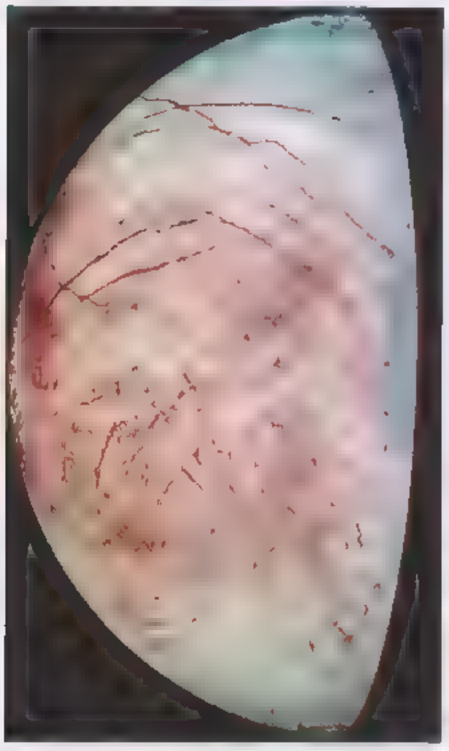


Figure 8.11 An image of Europa with the colours enhanced to show extra detail.

Its density (2990 kg m^{-3}) indicates that most of its interior must be rocky, but it is covered by bright ice to a depth of probably tens of kilometres. The surface has very few craters and is extremely flat.

- Using your answer to Activity 8.2, why would you deduce that the surface of Europa is very young?

  There are almost no craters on the surface, indicating that the surface material (water ice) has been solid and exposed to bombardment for a relatively short time.

Mottled, reddish 'chaotic terrain' exists where the surface has been disrupted and ice blocks have moved around (see Figure 8.12). The paucity of craters, together with these other features, has led scientists to conclude that there could be an ocean of liquid water beneath Europa's surface. Heat from radioactive decay in Europa's interior, supplemented by tidal heating, provides the energy to keep the ice molten at depth. The red material at the ridges and chaotic terrain is a non-ice contaminant and could be salts brought up from this possible ocean beneath the frozen surface. Figure 8.12 shows where the icy surface has been broken up into ice floes that floated apart until the water between them froze.

How do tidal forces produce heating of the interior of Europa? It is a result of deformation of Europa caused by Jupiter's gravity. Varying deformations can cause heating as exemplified by a squash ball, which is heated when it is repeatedly deformed by being struck. To understand how Jupiter causes the varying deformation of Europa that leads to the tidal heating, consider a very large, spongy ball falling towards Jupiter because of the gravitational attraction between the ball and Jupiter. Figure 8.13a shows the situation, where you can see that the ball really is *very* large! The ball can be notionally (not actually) divided into pieces of equal mass. This can be done in any way – Figure 8.13a shows two pieces of equal mass shown as red dots, one on the side nearest to Jupiter, and one on the side furthest from Jupiter. The force of gravity between two objects decreases as the distance between them increases.

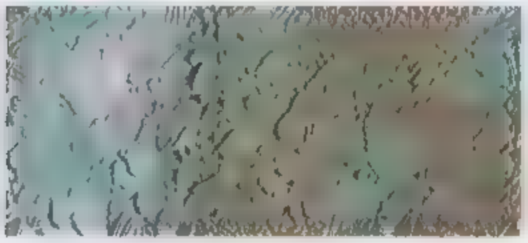


Figure 8.12 Close-up image of part of Europa, with enhanced colours, showing blocks of ice that seem to have drifted before being frozen in their current positions.

- On which of the two red dots will the force of Jupiter's gravity be greater?

  It will be greater on the dot nearer to Jupiter.

This will result in the two dots being pulled apart. The ball as a whole is distorted as shown in Figure 8.13b, the shape resembling that of a rugby ball (or an American football). Because the distortion arises from a difference in gravitational force across the ball, it is called a *tidal distortion*, and the bulges are called *tidal bulges*. If the ball has motion perpendicular to Jupiter it will orbit the planet (as seen in Section 1.3). However, this does *change* the tidal distortion. Suppose that the ball always keeps the same face towards Jupiter, just as the Moon keeps the same face towards the Earth. In this case, the tidal bulges will always lie on the line from the ball to Jupiter, as in Figure 8.13c. The distortion, from the point of view of the ball's interior, is fixed, so there is no tidal heating. In fact the effect of tidal forces causes the rotation periods of large moons to match their orbital periods; this phenomenon is called **synchronous rotation**.

To understand tidal heating, consider the ball to be in an eccentric orbit around Jupiter, as in Figure 8.13d. For the moment, suppose that the tidal bulges always lie on the line from Europa to Jupiter. You can see that there is a smaller tidal distortion the further the ball is from Jupiter. This is because not only does the gravitational force decrease with distance but so does the difference in gravitational force across an object.

■ If the ball is at an infinite distance from Jupiter, what is the difference in the gravitational force of Jupiter across the ball?

□ The gravitational force due to Jupiter is zero at all points on the ball, so the difference is also zero.

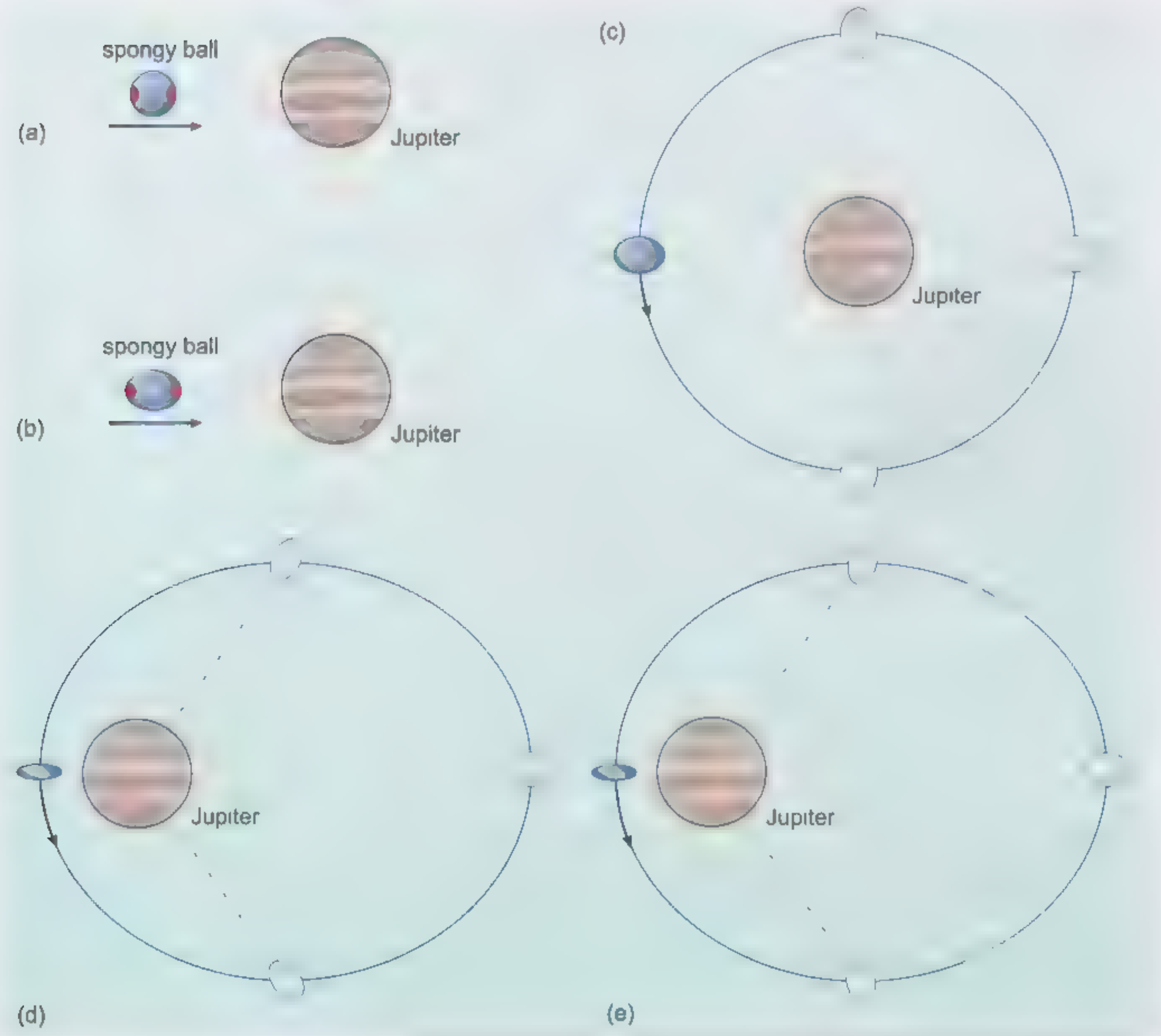


Figure 8.13 (a) A large, spongy ball falling towards Jupiter. (b) The distortion that would occur in the ball. (c) The ball in a circular orbit around Jupiter, and keeping the same face towards Jupiter. (d) The ball in a very elliptical orbit around Jupiter with the tidal bulges lying on the line from Jupiter (e) As (d), but with the tidal bulges failing to keep alignment. The effects are exaggerated to make them visible.

This variation in tidal distortion as the ball goes around its orbit is rather like cycling the degree of distortion of a squash ball, and so there is tidal heating. There is another contribution to the tidal heating because the tidal bulges on the ball do not stay in line with the direction to the planet This is a result of the combination of two different effects. The rotation of the moon tends to pull the tidal bulges out of alignment because of frictional forces. In addition, in an elliptical orbit, the orbital speed is not constant, but is greatest when the ball is closest to Jupiter. The direction to Jupiter therefore changes for any given position on the surface of the ball, which spins at a constant rate. Europa is sufficiently close to Jupiter for tidal deformation to provide the additional energy required to keep water liquid below the surface (see Figure 8.14). Further evidence for an ocean has come from the way in which Europa interacts with Jupiter's magnetic field, which suggests the presence of a subsurface conductive layer. This conductive layer could be provided by a salty liquid water ocean. In that widespread ocean aquatic life-forms could exist today. It would be an extremely ambitious mission to send a spacecraft to land on Europa and drill through the ice to make direct measurements, but missions to Europa are a high priority in the future plans of both the European Space Agency and NASA.

8.3.3 Other gas-giant satellites

Europa has an elliptical orbit around Jupiter so these effects can cause tidal heating You might wonder whether tidal heating can make other satellites of the gas giants into potential habitats. However, the temperatures of small satellites cannot be raised much by tidal heating or by any other form of heating. The reason is that the smaller a body, the greater its surface area per unit of its mass, thus leading to more rapid loss of heat by radiation to space. In order of distance from Jupiter, the large 'Galilean' satellites (they were discovered by Galileo Galilei in 1610 when Jupiter was first observed through a telescope) are Io, Europa, Ganymede, and Callisto. Io is thus closer to Jupiter than Europa, and it is heated even more strongly. It is heated so much it has a highly volcanically active surface devoid of water. Ganymede and Callisto are further away and the tidal forces are consequently smaller. However, astronomers believe both Ganymede and Callisto may also have subsurface oceans of electrically conducting salt water. This hypothesis is based on measurements of the moons' effects on Jupiter's magnetic field (Ganymede is the only moon that has its own magnetic field, produced by its liquid, iron rich, core). Since tidal forces alone are insufficient to melt the ices, melting is believed to be the result of trapped heat from the decay of radioactive materials in the rocky cores of these bodies.

The large satellites of Saturn, Uranus and Neptune are also insufficiently heated by tidal forces for melting to occur although subsurface oceans cannot be ruled out for the largest satellites. It was therefore a surprise when the Cassini spacecraft revealed a plume of gas and icy particles being ejected from near the south pole of Saturn's 500 km diameter moon Enceladus (Figure 8.15). It orbits in the densest part of the E-ring, one of the faint rings beyond those visible in small telescopes and this cryovolcanic activity provides the source for particles in the ring. The elevated temperatures near the plume sources, impurities in the plume particles directly measured by Cassini, together with models of the ejection mechanism all provide evidence for a liquid water reservoir below the surface. Although Enceladus experiences episodic tidal heating resulting from orbital interactions with another satellite of Saturn this appears to be insufficient to melt the ice and the heating mechanism is not currently understood.

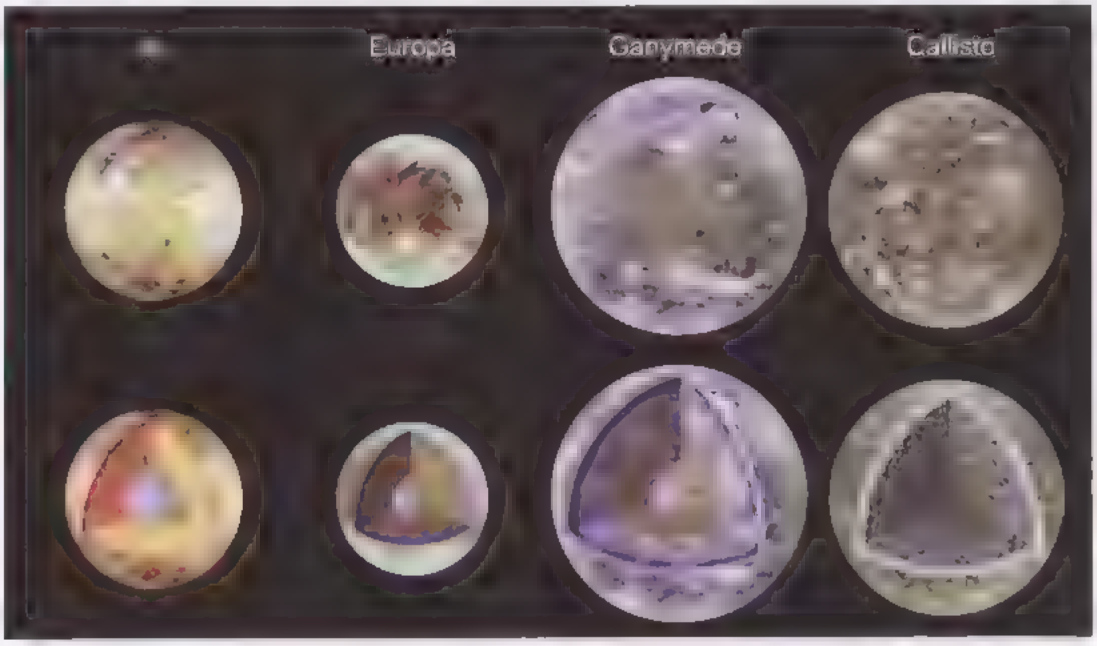


Figure 8.14 The four Galilean satellites of Jupiter with cutaway views of their possible interior structure. All have rocky or iron-rich cores, except Callisto. Io is characterised by active volcanoes whereas the other three moons have ice-rich crusts and possible subsurface oceans.

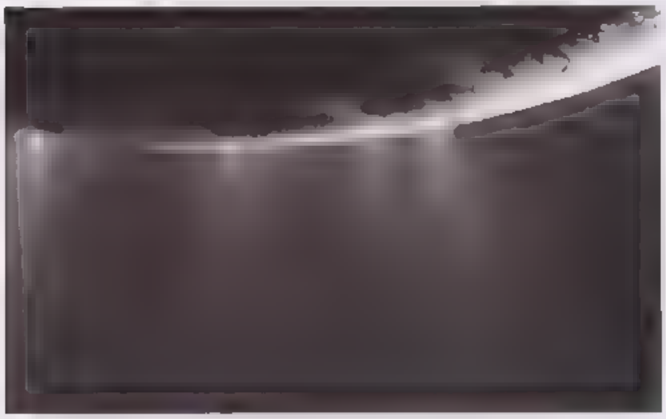


Figure 8.15 Plumes of icy particles, water vapour and organic compounds emanating from cracks in the crust near the south pole of Saturn's satellite Enceladus.

Saturn's largest moon Titan (Figure 8.16) is also of interest to astrobiologists for another reason. It is larger than the planet Mercury and the only planetary satellite with a substantial atmosphere. The atmosphere, composed primarily of nitrogen, contains a complex mixture of hydrocarbons formed from the break-up of methane by solar ultraviolet radiation. The atmosphere prevented views of the surface until the radar observations by Cassini and the landing of the Huygens probe in 2005. The conditions at the surface of Titan (1.5 times Earth's atmospheric pressure and a temperature of −180 °C) allow methane to exist in solid, liquid and gaseous form, in much the same way as water does on the Earth. Titan has been likened to a primordial Earth in deep freeze. The climate creates surface features similar to those found on Earth, such as rivers, shorelines, lakes (Figures 8.16) and sand dunes, but with methane and ice taking the place of water and rock. Titan may also have a subsurface ocean and it has even been suggested that a form of life using liquid methane rather than water might exist. Whether or not such exotic life could exist, Titan provides a fascinating and unique environment to study the type of pre-biotic chemistry that may have occurred on the early Earth and future missions are already being planned.

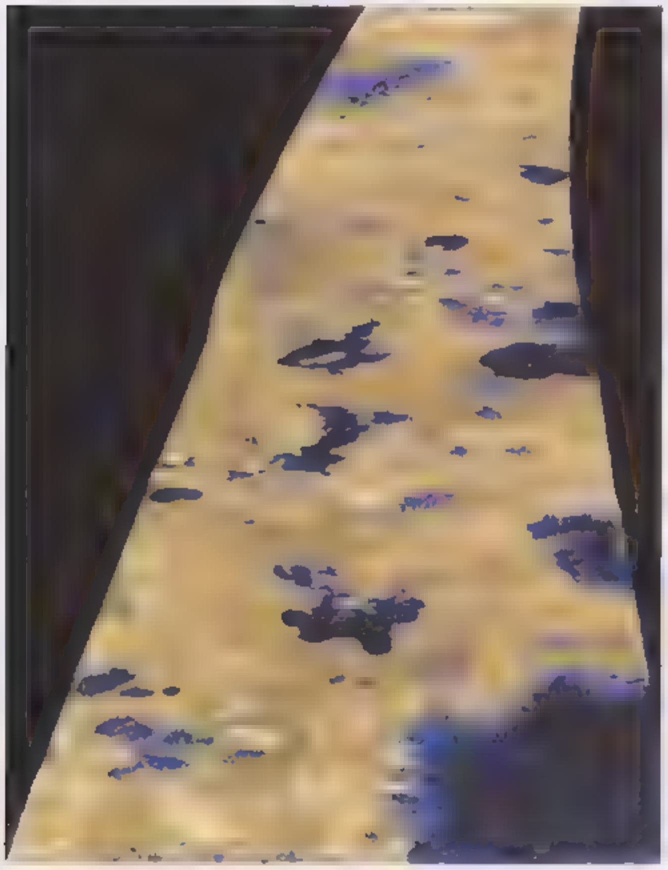


Figure 8.16 An image of a 140 km wide strip of the surface of Titan showing hydrocarbon lakes. The image was produced using the radar instrument on the Cassini spacecraft and the colours (which are false) represent the different strengths of radar signal returned.

8.4 Life in other planetary systems?

In the search for carbon water life, the type of potential extraterrestrial habitat that attracts almost exclusive attention is the surface regions of planets and their satellites, rather than interstellar clouds (such as those seen in Figures 6.11 and 6.12 and described in Section 6.7) and the surfaces of stars. In the case of stars, the high temperatures make them totally unsuitable for life. The reason for ignoring interstellar clouds is less obvious. It is because they are of such low density that it would be difficult to bring together enough atoms to build up huge carbon compounds in any quantity (although quite large molecules do appear to form), and because the pressures are far too low for liquid water.

In Section 7.6 you have already seen how difficult it is just to detect a giant planet orbiting close to a star.

- If planets are easier to detect when they are close to a star why can't we search for life on Earth-like planets in such orbits?

- We may be able to find a terrestrial planet (a rocky planet with mass similar to that of the Earth) orbiting close to a star but it will be too hot for life to exist.

Before searching for signs of life we need to determine which planets are most likely to satisfy the requirements for life described in Section 8.2.

8.4.1 Stellar environments

As you have determined from Activity 8.1, planets much closer to the Sun than the Earth (i.e. Mercury and Venus) do not satisfy the conditions for life. For a star like the Sun, the range of distances at which water can be in liquid form on the surface of a planet is relatively close to 1 AU. This range of distances is called the **habitable zone**.

The potential for life to develop also depends on the properties of the planet, such as its orbital eccentricity, obliquity or rotation period. A planet with a high eccentricity will experience higher average temperatures when it is close to the star. If its orbit takes it outside the habitable zone then it may experience extremes of temperature too great for life to develop or survive. A planet with a slow rotation (a long day and night) can undergo large extremes of temperature. In the Solar System Mercury has an average temperature of about 170 °C, but with a rotation period of nearly 60 days, it can reach 430 °C near midday and −160 °C at night. Also, if the obliquity is close to 90°, each pole of the planet will be exposed to sunlight for a half a year and then suffer half a year of night. This will also result in huge extremes of temperature.

The structure and composition of an atmosphere can also affect the surface temperature. An atmosphere helps reduce the range of temperatures between day and night and can also cause the average temperature to be higher due to the **greenhouse effect**. Certain gases, such as carbon dioxide and methane are transparent to visible light (from the star), which heats the planetary surface. The surface, at a low temperature compared to a star, radiates most of its energy at infrared wavelengths (see Section 2.3), which are absorbed by these gases, heating the atmosphere. The Earth benefits from a mild greenhouse effect (although there is currently concern over the increase in carbon dioxide content in the atmosphere and the potential consequences for the Earth's climate). Venus, with a very dense carbon dioxide atmosphere suffers temperatures of over 450 °C at its surface due to its very strong greenhouse effect. Highly reflective clouds or planetary surfaces can result in cooler temperatures because a fraction of the sunlight will be reflected directly back into space.

However, the dominant factor in determining the habitable zone for an exoplanetary system is the luminosity of the star.

- How will the habitable zone of an M class main-sequence star differ from that of the Sun?

□ The M class star has a much lower luminosity (Section 5.3) so its habitable zone will be much closer to the star.

Figure 8.17 illustrates the approximate distance of habitable zones about main-sequence stars of different spectral types. The more luminous the star, the further away will be its habitable zone.

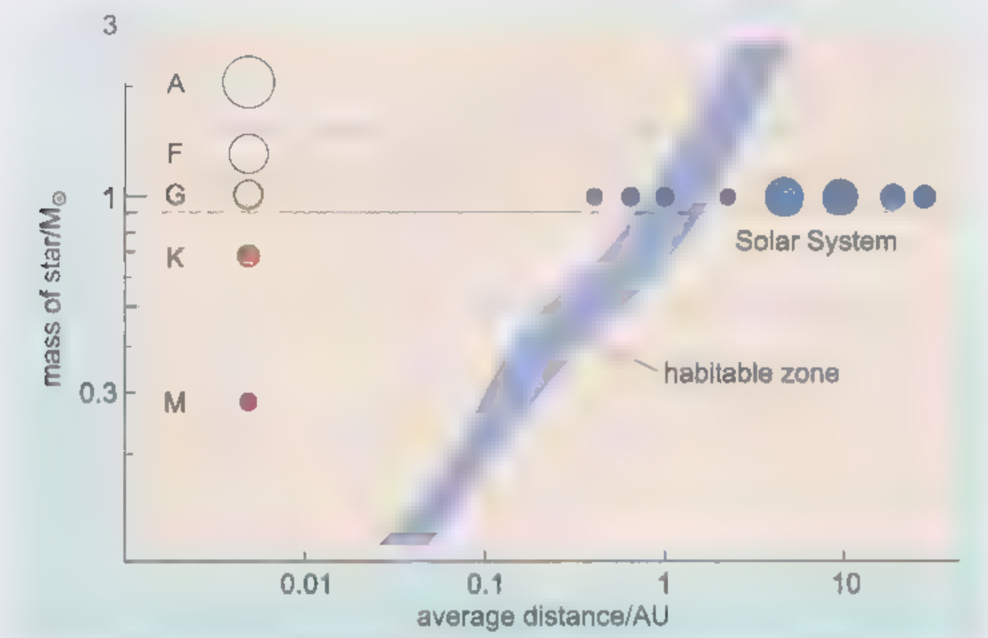


Figure 8.17 The location of habitable zones around different main-sequence stars compared with the Solar System (planet sizes not to scale).

■ Variable stars change in luminosity on timescales from minutes to months. Why does this prevent the establishment of a stable region where life could exist?

If the luminosity of the star changes, then the position of any habitable zone will also change on the same timescale. There may be no region where conditions suitable for life can exist throughout these changes.

Any change in luminosity of a star will result in a change in the position of its habitable zone. We know that stars change in luminosity as they evolve (Chapter 6). Stars like the Sun have relatively constant luminosity for about 10 billion years while they are burning hydrogen and lie on the main sequence of the HR diagram. Their habitable zones will therefore also be stable.

■ What will happen to the Sun after the end of its main-sequence life and how might this affect the location of its habitable zone?

□ The Sun will evolve into a red giant, with a higher luminosity. Its habitable zone will move outwards and the Earth will become too hot for life to survive.

The habitable zone around such a red giant star will be at a much greater distance from the star. It is possible that conditions on the surfaces of satellites of the gas giants in the Solar System, and in particular Titan, which already has a dense atmosphere, could become suitable for life to develop. However, the late stages of stellar evolution happen relatively rapidly compared with the time spent on the main sequence and the stars undergo huge changes in temperature and luminosity. There may therefore not be sufficient time for life to evolve beyond very primitive single-cell bacteria if the history of life on Earth is typical of life elsewhere.

- Why would a habitable zone around a high-mass main-sequence star of spectral type O or B not be a good location for life to develop?

- The lifetimes of high-mass stars are very short (Section 5.4) so life may not have time to develop on a planet in the habitable zone before the star ends its main-sequence life and changes luminosity or even engulfs the planet if it expands to become a red supergiant.

So, if the development of life on Earth is typical, a star must have a stable (i.e. main sequence) lifetime long enough for this development to take place. If we wish to detect the presence of life on another planet, this time must be long enough for life to have changed the atmosphere, through production of oxygen by photosynthetic plants (see Section 8.4.2). This requires stars of spectral types F, G, K or M. The presence of other stars or planets can also affect the potential for life to develop. About a third of the stars in our galaxy are in binary or multiple star systems. A planet in such a system must have a stable orbit within a habitable zone. This can occur if the stars in the binary system are sufficiently well separated or possibly if they are very close (see Figure 8.18). In both cases the orbit of the planet must not be perturbed too much by the changing gravitational forces of the two stars or be exposed to large changes in temperature. Similar limitations can apply for a terrestrial planet in a system with one or more massive giant planets, but in this case there is only the requirement for orbital stability since the planets are not luminous.

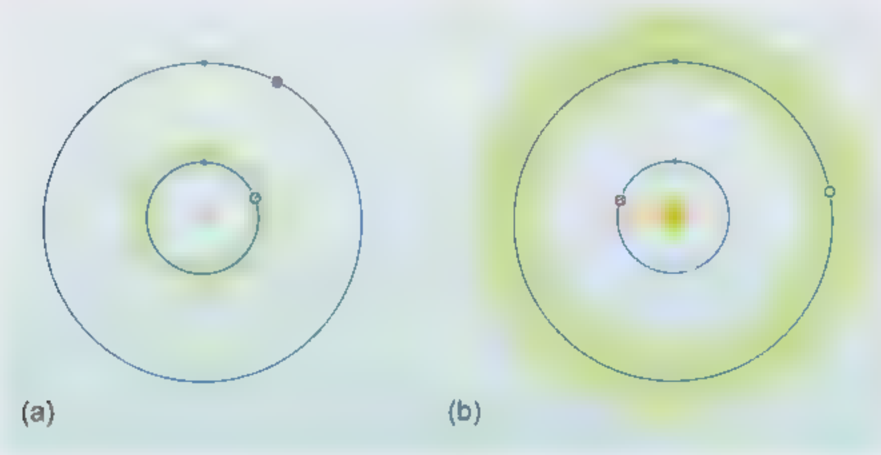


Figure 8.18 Planets in habitable zones in binary star systems. (a) The planet orbits within the habitable zone of one star, with the second star sufficiently distant not to perturb its orbit or significantly heat its surface. (b) The planet orbits both stars (circumbinary orbit) and is distant enough for the change in positions of the stars in their orbit not to cause large variations in surface temperature.

As you can see there are many factors (and they have not all been covered here) that can affect the likelihood of life forming and surviving. The next section will consider how it might be possible to detect life on other planets.

8.4.2 Searching for life on exoplanets

The only way of telling whether a planet is inhabited is from the electromagnetic radiation it produces. Section 8.5 considers radiation that may be intentionally produced by intelligent life, but it is possible to detect the environments in which life may have developed, or the effects of such life, using the spectrum of a planet. The basic sorts of spectra (continuous, absorption and emission) were introduced in Section 2.2. Consider first what could be learned from the *continuous spectrum* of a planet.

Examples of continuous spectra are shown in Figure 2.7. The spectrum of the Earth has two components. One has almost the same shape as that of the Sun (it is sunlight reflected from the Earth) and the other has broadly the same shape but at far longer wavelengths (see Figure 8.19). This second component of the Earth's spectrum is concentrated at infrared wavelengths, which are emitted by the surface and the atmosphere. If you had infrared eyes the Earth's surface and the atmosphere would seem to glow!

■ What do the wavelengths of emission from the Earth tell you about the Earth's average surface temperature, compared with the objects whose spectra are depicted in Figure 2.7?

□ Comparison with Figure 2.7 shows that the Earth's peak emission is at a much longer wavelength so its surface must be a lot cooler than 3000 K!

In fact, the Earth's average surface temperature can be inferred to be about 15 °C. From other information in the spectrum (beyond the scope of this module) the surface pressure can also be estimated, and from a vantage point in space, the pressure and temperature at the Earth's surface could be proved to be suitable for water to be liquid. But would it be possible to tell whether water is present? The answer is yes, from the absorption features in the spectrum.

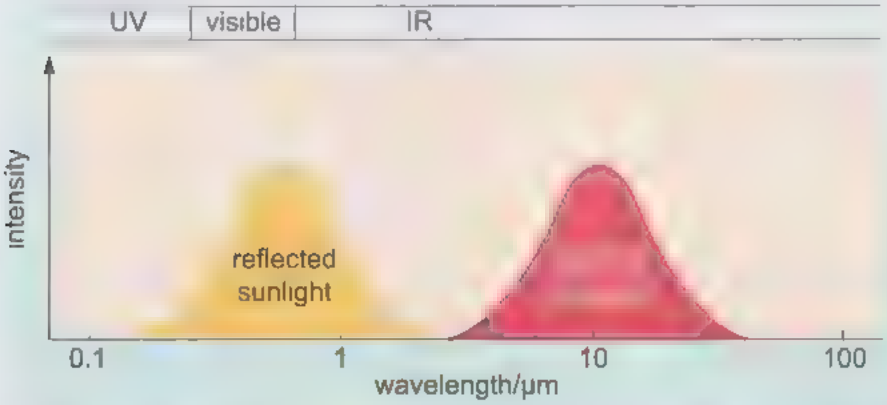


Figure 8.19 The continuum spectrum of a planet like the Earth.

Figure 2.2b shows an absorption spectrum, where material in an absorbing cloud has depleted the light at certain wavelengths From the wavelengths of these absorption lines it is possible, in principle, to determine the composition of the cloud. Likewise, the spectrum of the Earth also shows absorption lines or bands from absorption in the Earth's atmosphere. The reflected component of the Earth's spectrum does contain some absorption lines, as well as features due to surface materials – in particular from vegetation, but the most prominent absorptions are found at the wavelengths of the emitted component Some of these absorption bands show that water (H_2O) is present as a gas (vapour) in the atmosphere (see Figure 8 20). It is even possible to tell that there is so much H_2O present that some of it must have condensed on the surface, where it would be predominantly in liquid form rather than solid ice Thus there is the first requirement for life – liquid water – and it follows that the average surface temperature is low enough for huge carbon compounds to be stable. The presence of carbon is indicated by other absorption lines such as those caused by carbon dioxide (CO_2) in the atmosphere. However, you should note that this alone does not indicate that complex molecules for life can survive, because Venus has very strong carbon dioxide bands but is hostile to life.

It's therefore possible for a distant alien astronomer to establish that the conditions for carbon–water life exist on the Earth, but not (yet) that life is actually present The clue to this would be the presence of O_2, the oxygen that we breathe. Oxygen is such a reactive substance that it would rapidly disappear from the atmosphere unless it was being regenerated almost continuously. The best way of accomplishing this is photosynthesis by green plants and other organisms. Unfortunately, O_2 does not have strong lines in the infrared spectrum. However, an alternative strong indication is an absorption line caused by ozone (O_3). Through chemical processes in the atmosphere this is derived from O_2. From the ozone line the alien astronomer would be able to establish that oxygen as O_2 is a major constituent of the Earth's atmosphere, so may infer the existence of life.

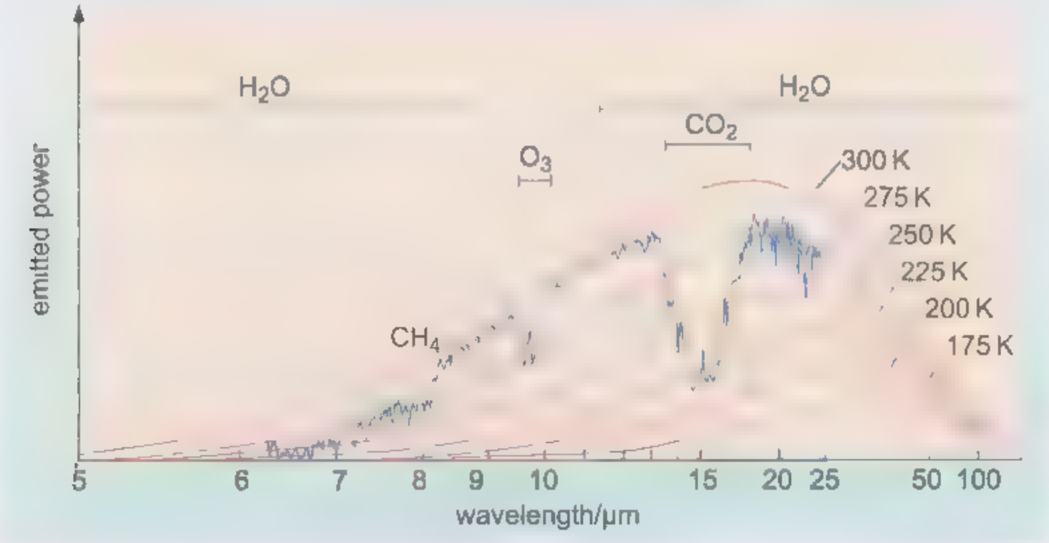


Figure 8.20 The emission spectrum of a planet like the Earth, showing absorption bands from some atmospheric gases.

In **photosynthesis** an organism starts building its body tissues, which include complex carbon compounds, from simple molecules, namely carbon dioxide and water. An energy source is required, and most types of organisms use solar radiation. Oxygen (O_2) is made as a by-product by most types of photosynthesising organisms. Without photosynthesis, the O_2 content of the Earth's atmosphere would be *far* lower. Thus the ozone absorption feature indicates strongly that the Earth is inhabited.

Directly measuring the spectra of Earth-like planets around other stars is not currently possible. There have been plans for space missions to attempt this, and although none are scheduled for launch at the time of writing, astronomers are confident that humans will be able not only to detect Earth-like planets but that we will have the capability to measure their spectra in the next few decades. We should then be able to establish whether liquid water exists at the surface and, therefore, whether the surface conditions are suitable for carbon–water life. If there is an ozone absorption feature we could be fairly confident that life is present. We would be even more confident if other absorption features were present, in particular, methane (CH_4), which is generated by both large organisms and bacteria. There are potential non-biological mechanisms that could produce each of the spectral signatures described above, but it is a combination of these features that would give confidence of a biological source.

Unfortunately, ozone could be below detectable limits, even if life is present, for any one or more of three reasons.

1 Local life-forms do not photosynthesise.

2 Local life-forms do photosynthesise but oxygen is not released (there are some terrestrial organisms for which this is the case).

3 The rate of release of oxygen is so slow that the atmospheric abundance, in the face of removal processes, remains low.

However, even if there were no ozone line, there could be other groups of absorption lines that would indicate the presence of life. Astrobiologists have identified such groups, but there is no space here for details.

8.5 Is there intelligent life out there?

8.5.1 How do we search for intelligent life?

Earlier on, this chapter examined the requirements for life and possible habitats in the Solar System and around other stars in which life may have developed. You have also looked at ways in which it may be possible to identify the planets on which life may reside. Such observations are extremely challenging and it has not yet been possible to detect such a planet. With the current knowledge of extrasolar planetary systems it appears likely that terrestrial planets are common, with many providing the conditions for life to develop. If the history of life on Earth is representative, then it is a long process for intelligent life to develop. If intelligent life was present on another planet in our galaxy, how could we detect it? The distances between stars are so vast that travel from one stellar system to another to find out would take an unfeasibly long time We would have to rely on detecting signs of intelligent life through observations of the electromagnetic radiation they transmit either accidentally or deliberately. This provides a convenient, if rather restricted, definition for 'intelligent' life, i.e. the ability to communicate signals across space. That communication may be deliberate (through transmission of signals with the express purpose of seeking contact) or accidental. For terrestrial-type life, this communication would almost certainly be at radio wavelengths, which are used for communication and can be transmitted through the Earth's atmosphere.

■ Humans have only been transmitting such radiation into space since the advent of radio communication a little over a century ago. What limitations would apply to an intelligent civilisation on another planet trying to detect our presence?

□ If radio signals started around 100 years ago, then they will have reached a distance of about 100 light-years from the Earth. An extraterrestrial civilisation would have to be within a distance of 100 light-years to be able to detect them.

There are a number of other problems in detecting such signals. The earliest signals would have been relatively weak and as they spread out though space they would have decreased in strength according to the inverse square law (see Section 1.5). However, since stars do not emit large amounts of electromagnetic radiation in the radio range (their spectra peak in the visible or near-infrared ranges), artificial signals could be distinguishable from natural sources with very sensitive receivers. Although most signals from the Earth were not intended for communications beyond the Solar System, their artificial nature would be easily recognised if they were detected, because of their complex but structured modulation over a narrow range of frequencies (in the radio region of the electromagnetic spectrum it is sometimes common to refer to frequencies, rather than wavelengths). Interpreting their meaning would probably be much more difficult. These factors also apply to any attempt to detect routine radio signals from other stars. Although we may not be limited to a distance of 100 light-years (since another intelligent civilisation may have been transmitting for much longer), accidental signals would be too weak for us to detect unless they came from planets belonging to the nearest stars. These signals can be masked by natural radio emission from astronomical sources or by terrestrial interference. The level of such interference is one factor in the choice of frequencies to search for extraterrestrial signals.

If an extraterrestrial civilisation had deliberately chosen to draw attention to themselves, then we presume they would have chosen frequencies for which the interference is minimised and which we may recognise as significant. The potential range of detection will be much higher because, as a deliberate attempt is being made to communicate, the maximum power will have been employed in a narrow range of frequencies, and possibly in a narrow beam towards our star.

There have been a number of searches for extraterrestrial intelligence (SETI) dating back almost as far as the beginning of radio communication. Some of you may have contributed in a small way through the SETI@home project that used the enormous spare capacity distributed in small home computers to process the vast amounts of data. There is not the space to describe these here, but they have one thing in common – none have so far been successful!

8.5.2 How do we communicate with intelligent life?

Communication with extraterrestrial intelligences (CETI) using radio transmissions involves most of the issues already discussed for SETI. The question of attempting to respond would certainly arise if an extraterrestrial signal was detected. Alternatively, we could choose to deliberately transmit signals to announce our presence and wait for a reply.

- What is the limiting distance for us to receive a reply from a civilisation that has detected our own signal?

  If our earliest radio signals were detectable and responded to as soon as they had been received, then the maximum total signal travel time is about 100 years. The signals would have to travel in both directions in this time so any civilisation would have to be within 50 light-years.

Clearly, in order to have any reasonable dialogue, we should choose to direct our own communications towards nearby stars. Even then, it would be a painfully slow process with an exchange of information taking years or decades.

- The ~~only~~ first attempt made ~~so far~~ to send a deliberate transmission was in 1974, when the world's largest radio transmitter at Arecibo in Puerto Rico was used to send a simple repeated message. The target was the globular cluster M13 at a distance of 25 000 light-years. What are the pros and cons of this choice?

  A globular cluster contains many thousands of stars so one transmission could be directed towards many stars at once, increasing the possibility that one may host an intelligent civilisation. Globular clusters are old, so there will have been more time for a civilisation to develop, compared with nearer, young open clusters. However the extreme age of globular clusters means the stars will have formed when there was a much lower abundance of heavy elements and therefore are less likely to have terrestrial planets. Any reply from M13 would not arrive for at least 50 000 years, and in any case, the relative positions of the Sun and cluster will have changed over this time, with their motion about the Galaxy, so the cluster will not even receive the message.

The Arecibo transmission was really a symbolic demonstration of what could be done. There are scientists who caution against making our presence known by deliberate transmissions, or by responding to any received signals. This reticence stems from evidence of the way that human civilisations have acted when confronting less technologically advanced cultures. A vastly more advanced alien intelligence may regard us in the same way as we see, for example, a colony of insects! Even if we chose to try to communicate and directed effort into transmissions to nearby solar-type stars, what possibility is there that there will be anyone to hear? We consider this question in the next section, which investigates the Drake equation, an attempt to quantify the number of advanced civilisations in our galaxy.

8.5.3 How many civilisations are there?

The Drake equation, which was devised in 1961 at a meeting which established SETI as a scientific discipline, defines the number of civilisations in our Milky Way galaxy whose electromagnetic emissions are detectable, N, as a function of a number of different factors:

$$N = (R \times f_{\mathrm{p}} \times n_{\mathrm{e}} \times f_{\mathrm{l}} \times f_{\mathrm{i}} \times f_{\mathrm{c}}) \times L$$

In the next activity you will identify these factors and estimate this number.

Activity 8.3 The Drake equation

The estimated time for this activity is 45 minutes.

In the Activities section on the module website you will find some video clips from an Open University/BBC production about the Drake equation.

(i) Watch Clip 1 and write down the meaning of each term in the Drake equation.

(ii) Use the values specified for each term to determine the estimate of the value of N made in 1961 by Frank Drake and colleagues.

(iii) Watch Clip 2 to check your answer and learn about the Fermi paradox.

(iv) One solution to the Fermi paradox is that N is very low. This could occur because one or more of the terms in the Drake equation are much lower than originally believed. However, none of them can be zero because $N \geq 1$ (we exist). Alternatively, it may be that N is greater than 1 and we haven't really looked hard enough yet. If L is typically very small, then civilisations would not have had time to colonise other parts of the Galaxy to provide direct evidence of their existence and we would have to rely on detection of signals. There have only been reasonably complete searches for signals from stars out to a distance of ~600 pc. If the original estimate of N is correct, approximately how far away will the nearest planet hosting a civilisation be? To get a crude estimate, we can assume that the Galaxy is a uniform flat disc of stars with a radius of 50 000 light-years, i.e. with an area of $\pi \times$ (50 000 light-years)^2. The average separation of stars within the disc will then be typically

$$\text{average separation} = \sqrt{\frac{\text{area of the Galactic disc}}{\text{number of civilisations in the Galaxy}}}$$

What does this number imply about the possible value of N ?

(v) Even if we could *prove* that $N = 1$, why does this not rule out other intelligent life in the Universe?

Our knowledge of the values of the terms in the Drake equation is still too uncertain for us to have a sufficiently accurate answer to know if we have a reasonable chance of success with SETI, or even if we are not alone in the Galaxy. Currently, the first three factors on the right-hand side of the equation are better determined than the next four to the right. The first term, R, is well defined and observations of exoplanetary systems are rapidly improving our knowledge of f_p and n_e. The remaining values are the subject of much debate and speculation although some observational evidence about the value of f_l may be found in the next few decades. The discoveries of organisms that can survive in ever more extreme environments has led some astronomers to speculate that life will indeed develop wherever it is possible. However, we have no evidence, except our own existence, to determine how likely it is that this life will develop into intelligent life and how long it may take. The understanding of the astronomical environment (based on the stability and lifetime of the host star and the probability of astronomical catastrophes such as asteroid impacts) does help identify where the most suitable environments may be found. There are even fewer constraints on whether an intelligent species will become able and willing to communicate, and how long a civilisation will remain able to do so. The question is likely to be more relevant to sociologists and environmental scientists, since this length of time would be partly determined by the species' ability not to destroy itself through conflict or exhaustion of natural resources.

The possibilities of the presence of extraterrestrial life and other intelligent civilisations elsewhere in the Universe are of fundamental interest that transcends scientific investigation. However, it is only through the success of the scientific method that humanity has reached its current understanding of the properties, origin and evolution of galaxies, stars and planets and the parts they have played in the processing of matter to form the elements and molecules that make life possible.

End-of-chapter questions

Question 8.1

It has been suggested that, at temperatures too high for complex carbon compounds, life might be based on silicon. In one or two sentences, make an 'educated guess' about why chemists think this is unlikely.

Question 8.2

If Europa were in its own orbit at its present distance from the Sun, rather than in orbit around Jupiter, why would it be removed from the list of possible habitats for life today?

Questions 8.3 to 8.6 may be found overleaf.

Question 8.3

As the Sun ages, its luminosity (power output) will gradually increase considerably Explain why the Earth will eventually become uninhabitable.

Question 8.4

If there were no life on Earth, explain one way in which the infrared spectrum of the Earth would be different.

Question 8.5

If a spectrum was obtained of a planet considerably colder than the Earth, how would the spectrum differ from the Earth's?

Question 8.6

The choice of frequencies is just one factor in designing a programme to search for extraterrestrial radio signals. What other factors do you think are important? (Remember that as for any scientific research activity, limitations of funding may constrain what is possible or realistic.)

Now go to the module website and do the remaining activities associated with Chapter 8.

Questions: answers and comments

Question 1.1

The Sun has a diameter of 1.4 million km which is 1400 000 km. The scale of the model is 1 cm to 5000 km. The diameter of the Sun in the model will therefore be 1400 000/5000 = 280 cm = 2.8 m. There is no fruit large enough to represent it in the model!

Question 1.2

The distance from the Sun to the Earth is 150 million km = 150 000 000 km = 150 000 000 000 m. Light travels at 300 million metres per second so it will take 150 000 000 000/300 000 000 = 500 seconds to travel from the Sun to the Earth. This is a tiny fraction of a year (which is over 30 million seconds) so a light-year is not an appropriate unit to measure distances within the Solar System.

Question 1.3

Star X is four times further away than star Y, so it will appear to be $4^2 = 4 \times 4 = 16$ times fainter than star Y.

Question 1.4

The composition of the crust is dominated by silicon and oxygen whereas the whole Earth has a smaller proportion of these elements indicating that the proportion of rocky material is lower overall. The proportion of iron is much higher in the whole Earth than in the crust, indicating that the iron is concentrated towards the centre.

Question 2.1

Diameter of Jupiter is 140 thousand km = 140 000 km = 140 000 000 m = $1.4 \times 100\,000\,000$ m = 1.4×10^8 m.

Distance of Jupiter from the Sun is 778 million km = 778 000 000 km = 778 000 000 000 m = $7.78 \times 100\,000\,000\,000$ m = 7.78×10^{11} m.

Question 2.2

The energy of a photon of wavelength λ is given by $\varepsilon = hc/\lambda$ where h is Planck's constant and c is the speed of light.

For the X-ray photon: the wavelength is 1 Å = 1×10^{-10} m, and the photon energy is
$\varepsilon = hc/\lambda = (6.63 \times 10^{-34}\ \text{J s}) \times (3 \times 10^8\ \text{m s}^{-1})/(1 \times 10^{-10}\ \text{m})$
$= 2.0 \times 10^{-15}$ J.

For the radio photon: the wavelength is 1 m and the photon energy is
$\varepsilon = hc/\lambda = (6.63 \times 10^{-34}\ \text{J s}) \times (3 \times 10^8\ \text{m s}^{-1})/(1\ \text{m}) = 2.0 \times 10^{-25}$ J.

The X-ray photon is therefore $2.0 \times 10^{\ 15}$ J/$2.0 \times 10^{\ 25}$ J = 10^{10} times more energetic than the radio photon.

Question 2.3

From Table 2.4, the B-class star has a temperature of about 20 000 K and the M-class star has a temperature of about 3000 K. From Figure 2.7, an object with temperature 12 000 K or more has a spectrum that peaks on the short wavelength side of the violet end of the visible spectrum and will appear to have a bluish–white colour. The B-class star will therefore be expected to appear bluish–white. An M-class star, with a temperature of about 3000 K has a spectrum that peaks on the long wavelength side of the red end of the visible spectrum and will be expected to have an orange–white colour. From Figure 2.7 one would expect the B-class star to be much brighter than the M-class star if they are the same size and observed at the same distance. For the M-class star to be brighter it must therefore be much larger than the B-class star. (In reality most M-class stars aren't bigger than B-class stars, but there are some exceptions. The reasons for this will be provided in Chapters 5 and 6.)

Question 2.4

The diameter of the sunspot is equal to that of the Earth, i.e. 12 756 km = 1.2756×10^4 km = 1.2756×10^7 m and it is at a distance of 150 million km = 1.5×10^8 km = 1.5×10^{11} m. Its angular size is given by:
angular size in degrees = 57° × (actual size/distance)
= $57° \times (1.2756 \times 10^7 \text{ m})/(1.5 \times 10^{11} \text{ m}) = 4.85 \times 10^{-3}$ degrees.

There are 3600 arcseconds in a degree so the angular size of the sunspot is $4.85 \times 10^{-3} \times 3600$ arcsec = 17 arcsec.

The angular size of the Sun is about 0.5° = 0.5 × 3600 = 1800 arcseconds The sunspot is therefore 17 1800 = 0.0094 times (about 0.01 times or 1% of) the angular size of the Sun.

Question 2.5

(i) The distance 34.135 ly has too many significant figures because it is not known to any greater accuracy than 3 ly. It should be written as 34 ± 3 ly.

(ii) Both the distance and its uncertainty have too many significant figures. It would be more correctly stated as 29 ± 6 pc.

(iii) The uncertainty and the distance are not matched. The uncertainty implies that the distance is known to an accuracy of about one-hundredth of an AU, whereas the value is only quoted to an accuracy of 1 AU. If the distance is precisely six astronomical units then, in this case, it should be quoted as 6.00 ± 0.01 AU.

Question 3.1

The dark matter would be transparent, so it would look like an empty glass. Also, because the dark matter hardly interacts at all with normal matter (except through gravity), it would immediately fall through the bottom of the glass without doing any damage to it.

Question 3.2

Galaxy (a) is elliptical. It has a smooth, oval shape and no sign of spiral arms or a central bulge. Galaxy (b) is a spiral. It has a central, bright concentration of light (the bulge) and clear spiral arms. Galaxy (c) is irregular. It has an irregular shape and irregular distribution of light.

Question 3.3

If we distributed all the galaxies uniformly throughout the entire Universe, but lined up all the elliptical galaxies so the longest axis of every elliptical galaxy pointed in the same direction, then this would be homogeneous but anisotropic. Other possible arrangements that are homogeneous but anisotropic also exist.

Question 3.4

In the picture, the space between the 'galaxies' is being stretched, which is correct, but the 'galaxies' themselves are also being stretched, which is not correct. In the real Universe, galaxies are gravitationally bound and do not expand with the expansion of the Universe. Similarly, we don't get taller as the Universe expands, because our heads are connected to our feet.

Question 3.5

Redshift is defined in Section 3.7 as the change in wavelength, divided by the original wavelength. Reading off the graph, the wavelength is about 980 m. The precise value is 983.1 nm. The change in wavelength is 983.1 nm – 121.6 nm = 861.5 nm. To find the redshift you need to divide this by the original wavelength of 121.6 nm, so the redshift is 861.5/121.6 = 7.085. Your answer should be about 7. (You would get the same answer if you measured wavelengths in some other unit, such as metres, instead of nm.) It turns out that the light from this quasar has taken 94% of the lifetime of the Universe to reach us!

Question 4.1

Let's say you're 25 years old. (You may or may not agree with this estimate, but my mother has insisted that she's 24 for at least the last 30 years.) Section 4.3 gives one light-year as 9×10^{15} metres. The distance the radio signal will have travelled will be 25 times that, or about 225×10^{15} metres, which is more usually written as 2.25×10^{17} metres.

Question 4.2

(a) One way to calculate the answer is as follows. Imagine that you have a box containing 14 protons and two neutrons – the 7 : 1 ratio mentioned in the question. If a nucleus of helium-4 is made from two protons and two neutrons, there will be 12 protons remaining in the box, each of which can be considered as a hydrogen nucleus. Therefore there are 12 hydrogen nuclei for every one helium-4 nucleus in the Universe.

(b) Taking the mass of a helium-4 nucleus to be 4 units, and that of a hydrogen nucleus to be 1 unit, the relative masses of the helium-4 and hydrogen in the box are 4 and 12, respectively. The fraction of the mass in the box due to helium-4 is therefore 4/(4 + 12) = 0.25 or 25%, and that due to hydrogen is 12/(4 + 12) = 0.75 or 75%. As noted above this is approximately the mass fraction of helium in the Universe as predicted by more detailed calculations, and is in close agreement with that which is observed.

Question 4.3

Dark energy makes up the largest contribution to the energy density of the Universe. It does not clump, as far as we can tell, and it acts to speed up the expansion of the Universe. Dark matter makes up the next largest contribution. It's not clear what dark matter particles are, but dark matter clumps gravitationally. The gravitational effect of dark matter, like all matter, acts to slow down the expansion of the Universe. Protons and neutrons make up the next largest contribution. Most of the protons and neutrons are not in visible matter, but a minority make up visible matter, which is the smallest contributor to the energy density of the Universe.

Question 4.4

The cosmic microwave background is the light from when the Universe was last opaque. The distant regions of the Universe that we see now as the cosmic microwave background sent their light towards us nearly 13.7 billion years ago. In the meantime, the Universe has been expanding. Therefore, these distant regions are now even more distant. In fact they are now 48 billion light-years away, despite the fact that the Universe is only 13.7 billion years old.

Question 5.1

It's not possible to make observations of the death of such stars because their predicted lifetime (80 billion years) is significantly longer than the current age of the Universe.

Question 5.2

A star twice the mass of the Sun will be brighter than the Sun (by a factor of 16). Since the same nuclear reactions occur in both stars, more nuclear reactions must be occurring in the higher-mass star to provide its energy output, so more neutrinos will be released.

Question 5.3

The more massive a star is, the more rapidly it burns its fuel, so the shorter its lifetime. Because of this, a very massive star is expected to live for only a few million years, and certainly for a much shorter time than the Sun. Also, as the mass of a star increases, the rate of energy production increases at a faster rate than the force of gravity. Above a certain point gravity can no longer hold a star together and it is likely that a star 1000 times as massive as the Sun is not stable.

Question 6.1

The correct six terms in order of their stage in the life cycle are: dense cloud, protostar. main-sequence star, red giant, planetary nebula, white dwarf.

Question 6.2

The Eagle nebula is a region composed of cold dust and gas within which young stars are born.

The Ant nebula is an example of a planetary nebula. It is formed when a star of similar mass to the Sun runs out of fuel and ejects its outer layers.

The Crab nebula is the result of the death of a star much more massive than the Sun in a supernova explosion.

Question 6.3

This are two possible reasons for this: (1) the lifetime of the protostar is much shorter than that of a main-sequence star, hence one is far more likely to observe a star in the latter phase and (2) protostars are born in dense clouds of dust and gas which absorb the visible light they radiate making them much harder to detect.

Question 6.4

Stars would no longer be able to produce the energy required to support themselves against the force of gravity and would therefore collapse in upon themselves. (The process would halt when temperatures had become hot enough for Helium atoms to fuse together and release energy).

Question 7.1

(a) Uranus has an obliquity close to 90°. The seasons will therefore be far more extreme than on the Earth. The Sun will remain almost overhead for polar regions in the summer.

(b) Mercury's obliquity is 0° so there will be equal periods of daylight and darkness and the maximum altitude of the Sun at any given location will not change seasonally in the way it does on Earth. However, Mercury will have seasons caused by changes in its distance from the Sun because its eccentricity is high. 'Summer' will occur in the northern and southern hemispheres at the same time.

Question 7.2

Terrestrial planets are similar in size to the Earth and much smaller than the giant planets. They orbit much closer to the Sun than the giant planets. They are composed mainly of rocky materials with relatively thin secondary atmospheres of heavy molecules.

Giant planets are composed of rock and ice cores with dense envelopes of hydrogen and helium. Terrestrial planets have at most two moons whereas giant planets have large numbers of moons and extensive ring systems.

Question 7.3

From examination of the images and captions used for Activity 7.2: impact craters have floors lower than the ground level around the crater; volcanic craters tend to have raised flanks, or are located at high altitudes so their floors are above the local ground level. Impact craters are often surrounded by a blanket of ejected material and may have central peaks from 'rebounded' material. Volcanic craters can be associated with channels or ridges from lava flow.

Question 7.4

Jupiter orbits the Sun at a distance of 5.2 AU so at its closest to the Earth it will be at a distance of 4.2 AU (the Earth orbits at a distance of 1 AU). Now $4.2\ \text{AU} = 4.2 \times 1.5 \times 10^{11}\ \text{m} = 6.3 \times 10^{11}\ \text{m}$. A distance of 10 parsecs $= 10 \times 3.09 \times 10^{16}\ \text{m} = 3.09 \times 10^{17}\ \text{m}$. The Jupiter-like planet is therefore $(3.09 \times 10^{17})/(6.3 \times 10^{11}\ \text{m}) = 4.9 \times 10^{5}$ times further away. Using the inverse square law, it will therefore be about $(4.9 \times 10^{5})^{2} = 2.4 \times 10^{11}$ times fainter.

Question 7.5

Table 7.5 Preferred properties for exoplanet detection methods (completed version of Table 7.4).

	Planet's mass	**Planet's size**	**Planet's orbital distance**	**Distance from Earth**	**Star's mass**	**Star's apparent brightness**
Direct imaging	–	large	large	small	low	low
Astrometric method	high		large	small	low	
Radial velocity method	high	–	small		low	high
Transit method		large	small			-

Question 8.1

To quote Section 8.2.1, 'no other chemical element comes anywhere near carbon in its ability to form the large and complex compounds that are necessary for life'. Life is believed to be based on huge, complex chemical compounds, and silicon is unlikely to form sufficiently huge, complex compounds.

Question 8.2

If Europa no longer orbited Jupiter, it would no longer be tidally heated, and so its oceans would freeze. (It would then resemble many of the other satellites of the outer planets.)

Question 8.3

A considerable increase in solar luminosity will cause an increase in the Earth's surface temperature to the point where it is too hot for liquid water and for huge, complex carbon compounds. The Earth will then be uninhabitable. (The surface of Mars might have a period of being inhabitable. This is because a warmer surface will lead to gases, notably carbon dioxide, being released from the surface, thus increasing the atmospheric pressure to the point where water can exist as a liquid.)

Question 8.4

With no life on Earth there would be no photosynthesis, hence little oxygen, hence little ozone, and so the ozone absorption line would, at most, be very weak. (There would be other changes too, but these are beyond the scope of this module.)

Question 8.5

There are at least three changes that you might have thought of, based on the information given in this module.

1. The peak in the spectrum would be at longer wavelengths.

2. Life would be unlikely on a cold world, so there would be no ozone absorption line.

3. A cold atmosphere would contain little water vapour, so water absorption lines would be weak.

Question 8.6

There are many considerations and trade-offs in designing an extraterrestrial search programme. Some are listed below you may have thought of others.

Size of receiver.

Duration of search.

Range of frequencies.

Choice of target stars.

The largest receivers will allow detection of the faintest signals, but will be expensive to build and there will be competition for time on existing facilities. A more cost effective approach may be to use smaller receivers, which will have a more limited range of detectability, but could be operated for much longer in a dedicated search. The longer the duration of a search the greater the number of stars hosting potential transmitting civilisations that can be observed. The search could be directed at a relatively small number of individual stars, such as nearby solar-type stars, scanning a wide range of frequencies or scan a large area of sky at one or a few selected frequencies.

Comments on activities

Additional activities (at the end of each chapter) can be found on the module website. Comments on these are also on the website.

Remember also that there were activities that sent you directly to the module website (Activities 2.1, 3.1, 7.1 and 7.2).

Activity 1.1

In the first part, you might have found that the table-tennis ball did not move exactly in a straight line. This might happen if the surface was not level, or if either the ball or the table were not completely smooth. If you inadvertently spun the ball, and there was some friction between the ball and table, that would also drive it into a curved path.

In the second part, the cork flies off in the direction it was heading, and falls in a curved path towards the Earth. It is being pulled downward by gravity. If there was no gravity it would move horizontally in a straight line.

Activity 2.2

Here are some typical results from this activity:

diameter of coin = 1.7 cm (UK 5 pence coin)

distance of coin = 182 cm

distance ÷ diameter = 107.058 82 on a calculator (near enough 107).

So, for something whose angular size is that of the Moon (half a degree):

distance = 107 × diameter

diameter of Moon = 3476 km.

Therefore, distance of Moon = 107 × 3476 km, i.e. distance of Moon = 371 932 km (near enough 372 000 km).

To eclipse the Moon, the distance of the coin needs to be roughly one hundred times its diameter. If you tried using a coin larger than 2 cm diameter, it would need to be more than 2 m from your eye and would not fit on a 2 m rod.

Our result is quite close to the accurately measured value of 384 500 km; yours may be closer, or not quite so close. If you got a value of a few hundred thousand kilometres, that is reasonable. If your value was very different, check back through your calculations to see whether you have made a mistake, and look again at your measurements.

Activity 5.1

There are no additional notes for this activity.

Activity 6.1

After doing this activity, you should be able to explain why generally just one pulse is observed for each revolution of a pulsar, despite the beam emerging in two directions; two pulses would be observed only if the beam emerged at right angles to the pulsar's rotation axis. This activity also illustrates how pulsars might remain unobserved – they can only be seen if their radiation happens to be beamed in our direction.

Activity 8.1

Our answer consists of Table 8.1 preceded by a paragraph of overall considerations. You probably organised your answer differently. The 'Comment(s)' column in the table is not expected as part of your answer.

The most relevant data are the average surface temperature and surface pressure, because these determine whether water can exist as a liquid. The information about atmospheric composition is not so relevant because it lists only the main components: for example, for the Earth, water vapour is not included because it is a minor component.

Table 8.1 For Activity 8.1.

Planet	Potential habitat? Reason(s)	Comment(s)
Mercury	No. No significant atmosphere. Very large range of temperatures.	The average surface temperature is too low, although this is a little misleading – the day/night extremes are very high/low
Venus	No. Average surface temperature is much too high.	Temperatures vary little across the Venusian surface, so there are no cool niches.
Mars	Unlikely. Average surface temperature is too low. Tenuous atmosphere – is it below 6.1 millibars?	In fact, at certain times, the temperature can exceed 1 °C, so atmospheric pressure is the issue.
Ceres	No. Too cold, no significant atmosphere.	
Jupiter	Insufficient information given. The pressure and temperature increase with depth into the atmosphere, but it is not possible to say whether there is a level where liquid water could exist.	In fact, there is an atmospheric level where the temperatures and pressures would allow liquid water, and water is present in the atmosphere. Deeper down it is too hot. However, life is unlikely to emerge at any level in this atmosphere.
Saturn	As for Jupiter	As for Jupiter
Uranus	As for Jupiter	As for Jupiter
Neptune	As for Jupiter	As for Jupiter
Pluto	No. Too cold.	At its huge distance from the Sun, no part of Pluto's surface reaches anywhere near 0 °C. Also, although there is a tenuous atmosphere it is seasonal and pressure is much less than 6.1 millibars.
Haumea, Makemake, Eris	As for Pluto	

The satellites of the planets should also be considered. The Moon is ruled out because of its very low surface pressure and extremes of temperature but, among the satellites of the gas giants, there is at least one promising candidate, as you will see.

Activity 8.2

Impact craters accumulate on a surface and are subsequently obliterated by any resurfacing. Therefore, if a planetary body has a heavily cratered terrain and a lightly cratered terrain, the heavily cratered terrain will be older.

To turn this into absolute ages, astronomers use the variously cratered terrains on the Moon, for which there are known absolute ages. These have been obtained by radiometric dating of rock samples from these terrains (the details are beyond the scope of this module). There are difficulties in applying lunar data to Mars and, consequently, the absolute ages of the Martian terrains are poorly known.

One way of demonstrating the principle is to prepare a smooth surface of fine sand and throw water droplets at it every few seconds. The number of pits formed by the droplets grows with every throw. The pits can be removed by resurfacing the sand.

There are many other possible demonstrations. One advantage of using water is that, as in the real case, none of each projectile (water drop) remains. In the real case, they vaporise on impact; in this demonstration, they slowly evaporate.

Activity 8.3

(i) R is the rate at which suitable stars are formed (i.e. the number of stars formed per year),

f_p is the fraction of those stars with planets,

n_E is the number of those planets per planetary system with environments suitable for life (i.e. in the habitable zone of the star),

f_l is the fraction of those planets in which life appears,

f_i is the fraction of life-bearing planets in which intelligent life arises,

f_c is the fraction of planets with intelligent life that will become technologically advanced and develop a desire to communicate across space,

L is the lifetime of the communicating species, i.e. the length of time they will continue to transmit detectable signals into space.

(ii) The initial value of N estimated in 1961 was
$N = 10\ \mathrm{yr}^{-1} \times 0.5 \times 2 \times 1 \times 0.5 \times 1 \times 10\,000\ \mathrm{yr} = 50\,000$.

(iii) The Fermi paradox can be stated as 'If life is not rare, i.e. there are thousands of intelligent civilisations in the Galaxy, why have we not detected any trace of them?'

(iv) The area of the Galactic disc is π (50 000 light-years)^2 = 7.9×10^9 light-year^2. If there are 50 000 civilisations in the Galaxy then there will be one civilisation in every

$$\frac{7.9\times10^9 \text{ light-year}^2}{50\,000} = 158\,000 \text{light-year}^2$$

The average distance between civilisations will then be

$$\sqrt{158\,000 \text{light-year}^2} = 400 \text{ light-years}$$

If stars have been searched out to a distance of ~600 pc then we might have expected to have found a positive signal. This implies that N is smaller than the initial value derived by Drake, but can still be relatively large and remain consistent with our non-detection.

(v) Even if $N = 1$, there may be intelligent life present in any one of the other galaxies in the Universe.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover: © Lincoln Harrison

Figure 1.1: Lynette Cook/Science Photo Library; Figure 1.3m: © MShep2/ iStockphoto.com; Figure 1.3Mm: NASA http://visibleearth.nasa.gov/; Figure 1.3Gm: Courtesy of STEREO (NASA); Figure 1.3Em: NASA, ESA and Nota, A. (ESA/STScI/AURA); Figure 1.3Zm: NASA, ESA and the Hubble SM4 ERO Team; Figures 1.3Ym & 1.15: Chris Mihos (Case Western Reserve University)/ESO; Figures 1.7 & 1.12: © Dave Jarvis, Reproduced under Creative Commons Share Alike 3.0 Unported License; Figure 1.8a: Courtesy of SOHO, ESA and NASA; Figure 1.8b: Big Bear Solar Observatory, Figure 1.9: Professor Miloslav Druckmuller; Figure 1.11: ESA; Figure 1.13: ESO/Guisard, S. (www.eso.org/~sguisard); Figure 1.14a: NASA, ESA and the Hubble Heritage (STScI AURA)-ESA Hubble Collaboration. West, M. (ESO, Chile); Figure 1.14b: European Space Agency & NASA; Figure 1.14c: ESO; Figure 2.2: Adapted from Kauffman, W. J. (1987) *Discovering the Universe*, W. H. Freeman & Co.; Figure 2.3: X-ray: NASA/ CXC Wesleyan Univ Kilgard, R. et al.; UV. NASA JPL-Caltech; Optical: NASA/ESA/Beckwith, S. & Hubble Heritage Team (STScI/AURA); IR: NASA JPL-Caltech U of Arizona Kennicu, R , Figure 2.4a: Image courtesy of ESA; Figure 2.4b: ESO Heyer, H. H.; Figure 2.4c: ESA/AOES Medialab, Figure 2.4d: VLA, NRAO, Figure 2.5: SOHO (ESA & NASA); Figure 2.6: ESA & NASA; Olszewski, E. (U. of Arizona) HST; Figure 2.8: National Optical Astronomy Observatory Association of Universities for Research in Astronomy National Science Foundation; Figure 2.10: Dennis di Cicco Sky Publishing; Figure 2.12a: NASA, ESA, AURA Caltech, Palomar Observatory; Figure 2.12b: NASA, JPL-Caltech, Martha Boyer (Univ of Minnesota), et al.; Figure 3.2: NASA/Jason Ware; Figure 3.3a: NASA/JPL-Caltech/Hurt, R. (SSC); Figure 3.3b: Paul Doherty; Figure 3 4: Alex Cherney www.terrastro.com; Figure 3.5a: NOAO/AURA/NSF, Twardy, B., Twardy, B. and Block, A. (NOAO); Figure 3.5b: Alar Toomre (MIT) and Juri Toomre (University of Colorado); Figure 3.7: NASA, Andrew Fruchter and the ERO Team [Sylvia Baggett (STScI), Richard Hook (ST-ECF), Zoltan Levay (STScI)] (STScI); Figure 3.9: The 2dF Galaxy Redshift Survey team http://www2.aao.gov.au/ 2dFGRS , Figure 3.10a left: A. Fujii; Figure 3.10a right: R. Williams (STScI), the Hubble Deep Field Team and NASA; Figure 3.10b: Robert Williams and the Hubble Deep Field Team (STScI) and NASA; Figure 3.11: Günter Kerschhuber, Gahberg Observatory; Figure 3 12: Adapted from Yates, M. G. and Garden R. P. (1989) 'Near-simultaneous optical and infrared spectrophotometry of active galaxies', *Monthly Notices of the Royal Astronomical Society*, **241**, 167, The Royal Astronomical Society, Figure 3.13: Adapted from http://adsabs.harvard.edu/abs/2008A%26A...487..837F, Franceschini, A., Rodighiero, M. and Vaccari, G. © ESO 2008; Figure 3.14: Courtesy of JAXA; Figure 3.15: Adapted from figure 4 from http://arxiv.org/abs/0912.4263; Figure 3.16: NASA ESA and the Hubble Heritage Team (STScI AURA); Figure 3.18: Adapted from Mortlock, D. J. et al. 2011, *Nature*, **474**, 616, http://www.nature.com/nature/journal/v474/n7353/fig_tab/nature10159_F1.html;

Figure 4.1a. NASA Goddard Space Flight Center COBE Science Working Group; Figure 4.1b: Courtesy of NASA/WMAP Science Team; Figure 4.3: Gunn, C. Not Knot Poster, Copyright © 1990 by The Geometry Center, University of Minnesota. All rights reserved; Figure 5.2: National Solar Observatory; Figure 5.5: ESO Espinoza, P.; Figure 6.1a: Credit: NASA, ESA, and the Hubble Heritage Team (STScI AURA)-ESA Hubble Collaboration: Whitmore, B. (Space Telescope Science Institute); Figure 6.1b: NASA CXC/SAO/DePasquale, J.; Figure 6.3: Corradi, R., Isaac Newton Group of Telescopes, NASA/CXC/SAO; Optical: NASA/STScI; Figure 6.5: NASA, ESA, and Hester, J. (Arizona State University); Figure 6.6: © Anglo–Australian Observatory; Figure 6.7: NASA CXC MIT UMass Amherst Stage, M. D. et al.; Figure 6.8 left: Kitt Peak National Observatory 0.9-meter telescope, National Optical Astronomy Observatories, courtesy of Bolte, M. (University of California, Santa Cruz); Figure 6.8 right: Harvey Richer (University of British Columbia, Vancouver, Canada) and NASA; Figure 6.9a: from *Universe* by Rowan-Robinson, M. published by Longman Group UK Ltd, 1990; Figure 6.10: Adapted from Pagel, B. E. J. (1997), *Nucleosynthesis and Chemical Evolution of Galaxies*, CUP, Cambridge; Figure 6.11: ESO; Figure 6.12: Taken from Smith, N. et al. (2000) 'Large-Scale Structure of the Carina Nebula', *The Astrophysical Journal*, **532**: L145 L148, 2000 April 1. © 2000, The American Astronomical Society, Figure 6.13 NASA, ESA and the Hubble SM4 ERO Team, Figure 6.15: NASA, Alan Watson (Universidad Nacional Autonoma de Mexico, Mexico), Karl Stapelfeldt (Jet Propulsion Laboratory), John Krist (Space Telescope Science Institute) and Chris Burrows (European Space Agency Space Telescope Science Institute); Figure 6.17: Wolfgang Brandner (JPL/IPAC), Eva K. Grebel (Univ. Washington), You-Hua Chu (Univ. Illinois Urbana-Champaign) and NASA; Figure 7.2a–c: NASA/JPL; Figure 7.2d: Adapted from Chodas, P. (NASA JPL); Figure 7.4a: ESA © 2007 MPS/DLR-PF/IDA; Figure 7.4b: NASA; Figure 7.5a: NASA/JPL/Space Science Institute; Figure 7.6: Credits: Montage by Emily Lakdawalla, Ida, Dactyl, Braille, Annefrank, Gaspra, Borrelly: NASA JPL/Ted Stryk; Steins: ESA/OSIRIS team; Eros: NASA/JHUAPL; Itokawa: ISAS/JAXA/Emily Lakdawalla; Figure 7.7a: NASA/JPL-Caltech/UMD/EPOXI Mission; Figure 7.7b © Akira Fujii, Figure 7.8: NASA, ESA and Kalas, P. (University of California, Berkeley, USA); Figure 8.1: Courtesy of Derek Martin; Figure 8.3: NOAA PMEL Vents Program; Figure 8.4: Professor Charles Cockell; Figures 8.5b & c: NASA, James Bell (Cornell Univ.), Michael Wolff (Space Science Inst.) and The Hubble Heritage Team (STScI/AURA); Figure 8.6: NASA/MGS; Figure 8.7: NASA/JPL/Malin Space Science Systems; Figures 8.8 & 8.9: ESA DLR FU Berlin (Neukum, G.); Figure 8.10: NASA; Figure 8.11: NASA JPL University of Arizona; Figure 8.12: Galileo Project/JPL/NASA; Figure 8.14: Schenck, P. (2010) *Atlas of the Galilean Satellites*, CUP, Cambridge, Figure 8.15: NASA JPL Space Science Institute; Figure 8.16: NASA/JPL/USGS.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publisher will be pleased to make the necessary arrangements at the first possible opportunity.

Module team

S177 is based in part on the Open University module S194 *Introducing Astronomy*. The S177 module team would particularly like to acknowledge the contributions of the S194 module team for their work, which has been built upon in *Galaxies, stars and planets*.

Chair

Stephen Serjeant

Authors

Simon Clark

Simon Green

Stephen Serjeant

Critical Reader

Andrew Norton

Curriculum Manager

Nick Adams

External Module Assessor

Professor Gordon Bromage, University of Central Lancashire

Production Team

Jenny Barden (*Media Project Manager*)

Tracy Bartlett (*Curriculum Assistant*)

Duncan Belk (*Library Services*)

Martin Chiverton (*Producer*)

Corinne Cole (*Media Assistant*)

Vee Fallon (*Licensing and Acquisitions Assistant*)

Tot Foster (*Producer*)

Chris Hough (*Media Developer*)

Richard Howes (*Media Assistant*)

Martin Keeling (*Licensing and Acquisitions Coordinator*)

Peter Twomey (*Editor*)

Index

Index entries and page numbers in **bold** refer to important terms. Page numbers in *italics* refer to entries which are mainly or wholly in tables or figures.

S177 Galaxies, stars and planets

This is a 12-hour-long colour enhanced exposure of the sky in Australia. As the Earth rotates, the positions of stars appear to move on the sky, sweeping out circles. The colours of the stars are determined by their temperatures, which in turn depend on their masses and ages as you will find in this book. The colours of galaxies evolve too, as the stars within them age and new stars are born.

© Lincoln Harrison.

ISBN 978-1-7800-7334-7

9 781780 073347